

# Nanopore- and AI-empowered microbial viability inference

Harika Ürel<sup>1,2,3,4</sup>, Sabrina Benassou<sup>5</sup>, Hanna Marti<sup>6</sup>, Tim Reska<sup>1,2,3,4</sup>, Ela Sauerborn<sup>1,2,3,4</sup>, Yuri Pinheiro Alves De Souza<sup>7</sup>, Albert Perlas<sup>1,2,3</sup>, Enrique Rayo<sup>6</sup>, Michael Biggel<sup>8</sup>, Stefan Kesselheim<sup>5</sup>, Nicole Borel<sup>6</sup>, Edward J. Martin<sup>9</sup>, Constanza B. Venegas<sup>10,7</sup>, Michael Schlöter<sup>10,7</sup>, Kathrin Schröder<sup>10</sup>, Jana Mittelstrass<sup>11</sup>, Simone Prospero<sup>11</sup>, James M. Ferguson<sup>12</sup>, and Lara Urban<sup>1,2,3,4,8,\*</sup>

<sup>1</sup>Helmholtz AI, Helmholtz Zentrum Muenchen, 85764 Neuherberg, Germany

<sup>2</sup>Computational Health Center, Helmholtz Zentrum Muenchen, 85764 Neuherberg, Germany

<sup>3</sup>Helmholtz Pioneer Campus, Helmholtz Zentrum Muenchen, 85764 Neuherberg, Germany

<sup>4</sup>Technical University of Munich, School of Life Sciences, 85354 Freising, Germany

<sup>5</sup>Jülich Supercomputing Center, Forschungszentrum Jülich, 52428 Jülich, Germany

<sup>6</sup>Institute of Veterinary Pathology, University of Zurich, 8057 Zurich, Switzerland

<sup>7</sup>Institute of Comparative Microbiome Analysis, Helmholtz Zentrum Muenchen, 85764 Neuherberg, Germany

<sup>8</sup>Institute for Food Safety and Hygiene, University of Zurich, 8057 Zurich, Switzerland

<sup>9</sup>Institute of Ecology and Evolution, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FL, UK

<sup>10</sup>Institute of Medical Microbiology, Immunology and Hygiene, Technical University of Munich, 81675 Munich, Germany

<sup>11</sup>Eidgenössische Forschungsanstalt für Wald, Schnee und Landschaft WSL, 8903 Birmensdorf, Switzerland

<sup>12</sup>Garvan Institute of Medical Research, Sydney, NSW 2010, Australia

\*Correspondence address. Lara Urban, Helmholtz AI, Helmholtz Zentrum Muenchen, 85764 Neuherberg, Germany. E-mail: [lara.h.urban@gmail.com](mailto:lara.h.urban@gmail.com)

## Abstract

**Background:** The ability to differentiate between viable and dead microorganisms in metagenomic data is crucial for various microbial inferences, ranging from assessing ecosystem functions of environmental microbiomes to inferring the virulence of potential pathogens from metagenomic analysis. Established viability-resolved genomic approaches are labor-intensive as well as biased and lacking in sensitivity.

**Results:** We here introduce a new fully computational framework that leverages nanopore sequencing technology to assess microbial viability directly from freely available nanopore signal data. Our approach utilizes deep neural networks to learn features from such raw nanopore signal data that can distinguish DNA from viable and dead microorganisms in a controlled experimental setting of UV-induced *Escherichia* cell death. The application of explainable artificial intelligence (AI) tools then allows us to pinpoint the signal patterns in the nanopore raw data that allow the model to make viability predictions at high accuracy. Using the model predictions as well as explainable AI, we show that our framework can be leveraged in a real-world application to estimate the viability of obligate intracellular *Chlamydia*, where traditional culture-based methods suffer from inherently high false-negative rates. This application shows that our viability model captures predictive patterns in the nanopore signal that can be utilized to predict viability across taxonomic boundaries. We finally show the limits of our model's generalizability through antibiotic exposure of a simple mock microbial community, where a new model specific to the killing method had to be trained to obtain accurate viability predictions.

**Conclusions:** While the potential of our computational framework's generalizability and applicability to metagenomic studies needs to be assessed in more detail, we here demonstrate for the first time the analysis of freely available nanopore signal data to infer the viability of microorganisms, with many potential applications in environmental, veterinary, and clinical settings.

## Introduction

While microbial cultivation remains a foundational technique in microbiology to assess the taxonomic composition of microbial communities and to understand their physiology and ecosystem functions [1], only a small fraction of microbial diversity has been isolated in pure culture [2]. This limitation has led to undiscovered functions and biased representations of the phylogenetic diversity of microbial communities in nearly all of Earth's environments [2]. While medically relevant microorganisms of the human microbiome often constitute an exemption since they have been disproportionately well studied through microbial cultures [3], the clinical application of microbial cultivation for pathogen profiling is further limited by its time-consuming and labor-intensive nature [4].

The first studies of the so-called “microbial dark matter” have been enabled by advances in culture-independent molecular methodology [5] and have been based on amplifications of conserved marker regions such as ribosomal RNA genes [6]. Such targeted metabarcoding approaches, however, suffer from several limitations: they can often not provide strain- or even species-level taxonomic resolution, are highly dependent on genomic database completeness, do not allow for any functional inferences or virulence annotations, and often introduce amplification bias due to differential amplification efficiency and primer mismatches, which can significantly distort the representation of microbial community compositions [7].

Metagenomics, on the other hand, is a shotgun sequencing-based molecular methodology that can assess the entirety of DNA

Received: September 11, 2024. Revised: June 30, 2025. Accepted: August 7, 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

isolated from an environment or a sample and *de novo* assemblies of potentially complete microbial genomes of all present microorganisms; such genome-based approaches provide a variety of phylogenetically informative sequences for taxonomic classification, information about the metabolic and virulence potential of microorganisms, and the potential to identify completely novel genes [8, 9].

Especially long-read metagenomic approaches have shown great promise in achieving highly contiguous *de novo* assemblies through the recovery of high-quality metagenome-assembled genomes (MAGs) from complex environments; specifically, the latest advances in nanopore sequencing technologies have resulted in high sequencing accuracies of very long sequencing reads of up to millions of bases, which have allowed for the generation of hundreds of MAGs from metagenomic data, including the generation of closed circularized genomes [10, 11]. Nanopore sequencing technology is based on the interpretation of the disruption of an ionic current due to a motor protein guiding individual nucleotide strands through nanopores embedded in an electrically resistant polymer membrane at a consistent translocation speed [12]. This raw nanopore signal, or “squiggle” data, can then be translated into nucleotide sequence using bespoke neural network-based basecalling algorithms [13], which—when efficiently embedded on powerful GPUs—can generate genomic data in real time. The portable character and straightforward implementation of nanopore sequencing at low upfront investment costs further make this technology accessible for fast microbial and pathogen assessments at the point of interest all around the world, including in low- and middle-income countries [14].

In contrast to cultivation-based approaches, molecular methods suffer from their inherent deficiency of not being able to differentiate between viable and dead microorganisms [2, 15]. While cultivation-based approaches only detect viable microorganisms, DNA might remain intact and therefore accessible by molecular methods despite the respective microorganisms being dead [15, 16]. This would be relevant in the context of clinical infection prevention and control and pathogen monitoring, where certain disinfection methods or the use of systemic antibiotics often kill the bacteria before the DNA is destroyed [15, 17], but also for understanding the ecosystem functions of thus far understudied microbiomes [2]. For example, the air microbiome has been shown to be remarkably diverse and variable when assessed through nanopore metagenomics [18], but given the low biomass of this environment, it is expected that many microorganisms might be dead and stem from adjacent environments such as soil or water. The persistence of the DNA of dead microorganisms in the environment might hereby depend on many factors, including external conditions, such as temperature, pH, and microbial activity, and internal, taxon-specific parameters, such as microbial cell wall composition. Viability-resolved metagenomics would, however, be crucial for the interpretation of metagenomic data, ranging from outbreak source detection [19], food safety [20], and public health investigations [21] to ecosystem function inferences [22].

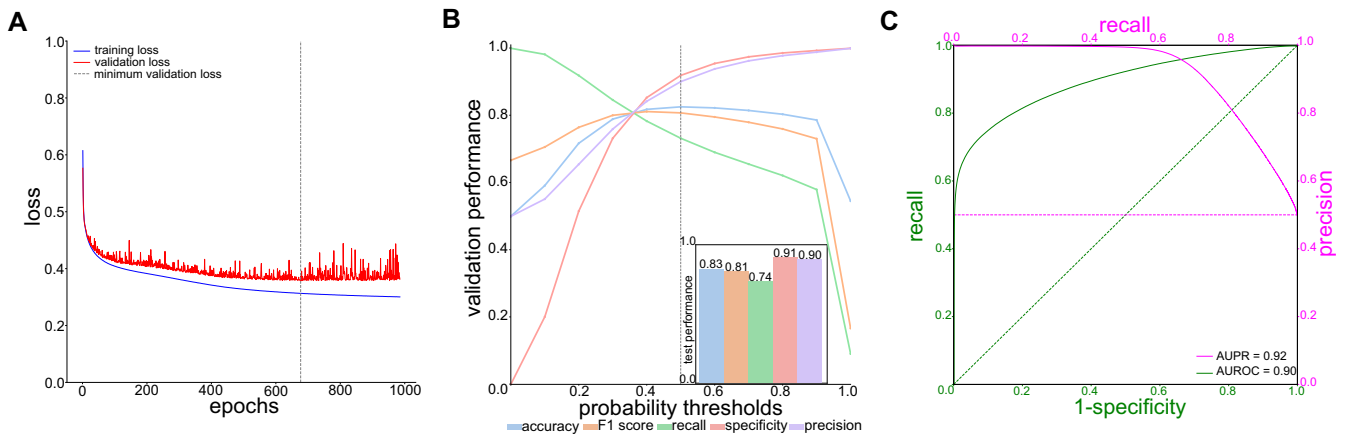
To assess microbial viability from genomic data, several approaches have been developed: culture-dependent viability methods combine the advantages of cultivation-based and molecular approaches by growing certain microorganisms of interest on selective media; this approach, however, remains time-consuming and labor-intensive and suffers from the same selectivity of growth media and culturable microorganisms as purely cultivation-based approaches [23], especially for fastidious or obligate intracellular microorganisms [24]. Microbial viability has fur-

ther been described by metabolic activity, where microbial cells are incubated with specific substrates, leading to ATP production, tetrazolium salt reduction, or radiolabeled substrate incorporation [25]. Further, ribosomal RNA may be assessed as a read-out of microbial activity [26]. To what extent such metabolic activity can be used as a proxy for microbial viability, however, remains to be explored [26]. While messenger RNA has been used as a viable/dead marker due to its intrinsic instability outside of the microbial cell [27, 28], the metatranscriptome still has to be stable enough in the environment to be detectable at all, potentially leading to many false-negative detections; if only 1 gene is targeted, the analyzed gene further has to be expressed shortly before cell death. Additional potential problems stem from the relatively challenging extraction protocols due to the RNA's instability and from the evolutionary conservation of gene sequences, which can hamper taxonomic resolution [15].

Finally, an aspect that can be used for viability-resolved metagenomics is the physical difference between viable and dead cells: viability PCR (vPCR) uses DNA-intercalating dyes such as ethidium monoazide (EMA) or propidium monoazide (PMA) to differentiate between viable and dead cells. These dyes penetrate only dead cells with compromised membranes and bind to their DNA via covalent bonds upon photoactivation, preventing them from being amplified during subsequent PCR [15, 17]; this approach has been applied to a diverse array of Gram-negative and Gram-positive bacteria, as well as to assess the effectiveness of disinfection and heat treatment [25]. It, however, relies on the assumption that membrane integrity is a reliable indicator of viability, which can lead to overestimation of viability if cells lose viability without immediate membrane compromise [29] and can be biased by the dye's variable permeability across different microbial cell wall structures [30, 31]. The dependence of the approach on photoactivation further means that turbid material might hamper the efficiency of the dye [32].

All these established viability-resolved metagenomic approaches are labor-intensive, require additional reagents and sample processing, and are often biased and lack sensitivity. We here hypothesized that the raw, freely available nanopore signal from metagenomic datasets might be leveraged to infer microbial viability, assuming that the native DNA from dead microorganisms accumulates detectable squiggle signatures due to, for example, external damage through UV, heat, or drought exposure; the lack of DNA repair mechanisms; or enzymatic degradation activity [33–35]. Such an analysis framework could be fully computational and utilize squiggle data that are automatically obtained with nanopore sequencing. While raw nanopore data are known to contain information about epigenetic modifications [36–38] and oxidative stress at specific human telomere sites [39], the applicability to assess microbial viability has not yet been tested.

In this study, we produced experimental nanopore sequencing data from viable and UV-killed *Escherichia coli* cultures to optimize deep neural networks to predict viability just from the nanopore squiggle signal. We then applied explainable artificial intelligence (XAI) tools, which allow us to identify the specific nanopore signal patterns in the input data that allow the model to deliver high-accuracy predictions as an output. We show that our computational framework can be leveraged in a real-world application to estimate the viability of obligate intracellular *Chlamydia suis*, pointing toward the applicability of our model across taxonomic boundaries, including to species with highly complex life cycles. We finally explore the limits of our model's generalizability through antibiotic exposure of a simple mock microbial community, where we had to train a new killing method-specific model



**Figure 1:** Training and viability inference on UV-killed *E. coli* of ResNet1. (A) Model loss for training and validation datasets across 1,000 epochs; the minimum validation loss of ResNet1 was reached at epoch 677. (B) Prediction probability threshold optimization on the validation dataset resulted in a probability threshold of 0.5 for obtaining maximum accuracy. *Inset:* Performance of ResNet1 on the test dataset (Materials and Methods). (C) Test dataset performance of ResNet1 in terms of precision–recall (PR; magenta) and receiver operating characteristic (ROC; green) curves and their respective areas under the curve (AUPR, AUROC).

to obtain accurate viability predictions. While the extent of our computational framework's generalizability needs to be assessed in more detail, we here demonstrate for the first time the potential of analyzing freely available nanopore signal data to infer the viability of microorganisms, with many applications in environmental, veterinary, and clinical settings.

## Results and Discussion

### Viability model training and inference

We generated controlled training data by nanopore sequencing native DNA of viable and dead *E. coli* (Materials and Methods). We killed *E. coli* cultures using different stressors to then isolate the extracellular DNA and expose it to natural degradation. We obtained only enough DNA for subsequent shotgun sequencing from the viable culture and from the culture killed through rapid UV exposure (viable: 212 ng/μL; UV: 5.46 ng/μL; heat shock: 0.03 ng/μL; bead beating: 0.67 ng/μL; Materials and Methods). We repeated this experiment and confirmed that rapid heat shock, as well as bead beating exposure, again resulted in very low DNA concentrations, suggesting quick and complete DNA degradation. We hypothesize that UV exposure is the only stressor in our study that simultaneously destroys bacterial cell walls and inactivates DNA-degrading enzymes. In contrast, heat shock at 120°C and bead beating might not uniformly degrade all enzymatic activity [40, 41], potentially allowing residual DNA-degrading enzymes to persist and contribute to the degradation of genomic material during subsequent natural exposure. We therefore created nanopore shotgun sequencing of the viable and the UV-exposed culture, which resulted in 2.92 Gbases (Gb; median read length of 2,476 b) and 2.69 Gb (median read length of 1,606 b) of sequencing output, respectively (Materials and Methods).

We then tested the implementation of different neural network architectures to predict the binary viability state from the raw nanopore data (0 = viable; 1 = dead after UV exposure; Materials and Methods). We processed the *E. coli* nanopore signal, or “squiggle,” data, cut it into altogether 3,181,600 signal chunks of 10k signals, and separated the chunks into balanced training (60%), validation (20%), and test (20%) sets along each original sequencing read to avoid that signal chunks from the same read would end up in the same dataset (Materials and Methods). These signal chunks

were then treated as 1-dimensional time-series signal data of consistent length. We trained the different model architectures using different learning rates (LRs) up to 1,000 epochs, assessing the models' performance based on training and validation loss after each epoch (Materials and Methods; [Supplementary Table S1](#); [Supplementary Fig. S1](#)). The loss plot of our best-performing model, a residual neural network with convolutional input layers (configuration Residual Neural Network [ResNet] 1; LR = 1e-4; [Supplementary Table S1](#); [Supplementary Fig. S1](#); Materials and Methods), shows minimal overfitting when the minimum validation loss is reached at epoch 667 (Fig. 1A). The other residual neural network architectures (ResNet2, ResNet3), on the other hand, resulted in overfitting to the training data at any LR, and the transformer architecture did not reach the minimum validation loss of ResNet1 ([Supplementary Fig. S1](#)). We next only focused on ResNet1 and optimized its probability threshold using the validation set; in order to obtain a high accuracy, we maintained the probability threshold at the default value of 0.5 (Fig. 1B), which resulted in a good final performance on the test data with an accuracy of 0.83 and a F1 score of 0.81 (Fig. 1B, inset) as well as area under the curve (AUC) values of 0.90 (area under the receiver operating characteristic curve [AUROC]) and 0.92 (area under the precision–recall curve [AUPR]; Fig. 1C).

We also trained the same residual neural network architecture ResNet1 on the basecalled nanopore data of viable and dead *E. coli* at a standardized chunk size of 800 b, which roughly corresponds to the signal chunk size of 10k signals (Materials and Methods). Independent of whether we only basecalled the canonical bases or used a N6-methyladenine (6 mA) modification-aware basecalling model (Materials and Methods), the model could not be trained to distinguish viable from dead data just from basecalled DNA sequence data ([Supplementary Table S1](#)). This shows that our model captures patterns in the squiggle data that go beyond the encoding of nucleotides and their known epigenetic modifications. While this was expected since we used the same *E. coli* culture with the same reference genome to create the viable and dead datasets, we can rule out that our squiggle-based model captured any random differences in DNA sequence context between the 2 datasets that might have occurred by chance.

We additionally obtained the performance of ResNet1 for different signal chunk sizes ([Supplementary Fig. S2](#);

Supplementary Table S1), and we found that viability prediction performance was possible from a minimum chunk size of approximately 5k but that performance further improved with increasing chunk size. This shows that larger signal chunks contain more information that can be used by our model to make more accurate per-chunk predictions despite a consequently reduced size of the training dataset. We here stick to our original model with a chunk size of 10k signals, which resulted in relatively good performance (Fig. 1; Supplementary Table S1).

We finally trained 2 logistic regression models using only the read length and translocation speed per sequencing read as the input feature, respectively, to compare the performance of our squiggle-based deep neural network with simple baseline models. Both models substantially underperformed in comparison to our model (Supplementary Table S1). The read length model performed similarly to a random classifier (accuracy of 0.5; Supplementary Table S1), showing that the read length distribution difference between the viable and dead sequencing datasets cannot be leveraged to infer viability in this dataset. We also argue that while a difference in read length distribution between viable and dead microbes might be expected in other settings, for example, after stronger or longer degradation exposure, such a difference would have to be substantial to allow for accurate viability classifications of each individual sequencing read. In mixed microbial communities, the read length distribution would further be confounded by the microbial composition and the respective genome sizes. The translocation speed model reached a slightly higher accuracy of 0.59 (Supplementary Table S1), which might be explained by UV-induced twists or kinks in the DNA backbone having a slight impact on the translocation of the sequencing read.

## Explainable AI application

We implemented Class Activation Maps (CAMs) as an XAI method [42] to identify the most important regions in the nanopore signal data that inform the model's viability classifications (Materials and Methods; Fig. 2A). We found that "dead" signal chunks exhibited discrete regions of increased CAM values ("CAM regions" defined at CAM values  $>0.8$ ; Fig. 2B for several true-positive classifications of the test dataset). To confirm the importance of these CAM regions for the model's final predictions, we applied consecutive masking of the regions with the highest CAM values within each nanopore signal chunk (Supplementary Fig. S3 for several examples); we observed that the prediction probability for being classified as "dead" decreased with increased masking of CAM-relevant regions, either by consecutively masking regions using a consistent mask size or by increasing the mask size (from 100 to 2k signals; Fig. 2C). This shows that the CAM application reliably pinpoints patterns in the nanopore signal that are predictive for our viability model.

We used the CAM regions to manually investigate the squiggle signals and found that many CAM regions of "dead" signal chunks included a sudden substantial drop in the nanopore signal. We therefore developed a simple algorithm that identifies such sudden drops and applied this XAI rule to our test dataset (Materials and Methods; Fig. 2D). We here classified any signal chunk with at least one sudden drop as "dead" and all others as "viable." While this simplified algorithm led to a drop in overall performance, we could still reach a relatively good overall accuracy of 0.68 (in comparison to 0.83 of the full model; Fig. 2E). While the XAI rule maintained performance in terms of specificity and precision, we observed a substantial drop in recall in comparison to the full model

(now 0.39 instead of 0.74). This shows that while the absence of a sudden drop in the nanopore signal data seems to reliably predict viability, not all "dead" signal chunks contain such a sudden drop. While this sudden-drop detection still seems to be at the core of our model's interpretability (when focusing on high-confidence true positive chunks at  $P > 0.99$ , the recall increased to 0.68), the model seems to additionally detect more subtle patterns in the nanopore signal data, which allow it to increase recall while maintaining specificity and precision.

Based on our previous experience with squiggle data analysis [43, 44], we hypothesize that the substantial sudden drops in nanopore signal might be caused by a twist or kink in the DNA backbone, for example, from 6 to 4 photoproduct pyrimidine dimers. The drop would then mark the event of a pore getting blocked due to such damage. Such a twist could also lead to a stalling signal if it impairs the motor protein from processing the DNA strand, which we indeed partially observed in our data (e.g., top signal chunk of Fig. 2B, D). While we here hypothesize that UV exposure might have caused such twists in the DNA backbone, we intend to explore the biological, chemical, and physical features detected by squiggle-based viability models in more detail in future nanopore-based microbial studies.

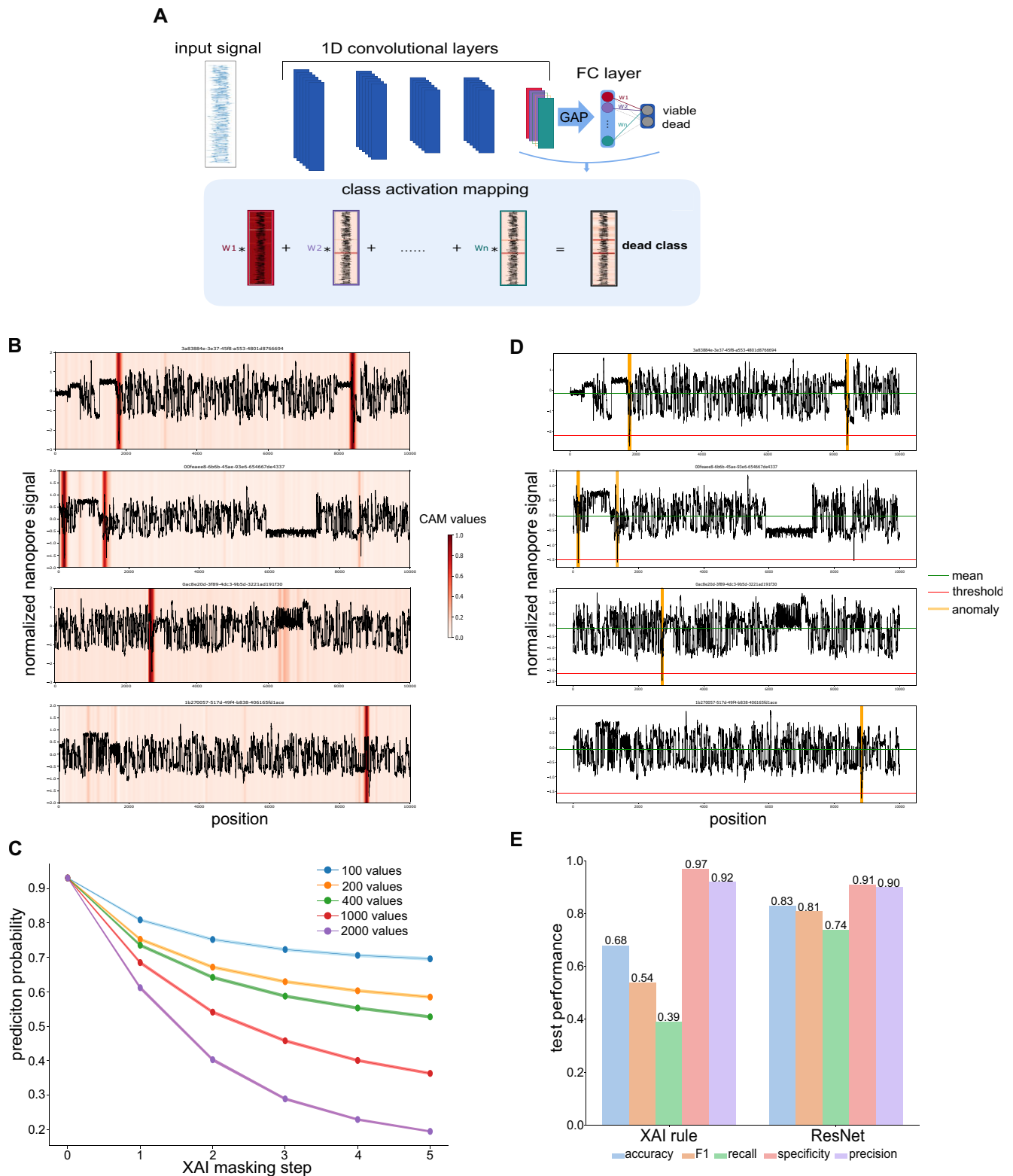
## Sequencing read-level viability predictions

We finally assessed the performance of our viability model on the sequencing read level instead of on the chunk level by leveraging the capability of ResNets to handle variable input lengths. When shifting the test dataset analysis from the chunk to the read level (Materials and Methods), the prediction performances increased substantially, from an accuracy of 0.83 (chunk) to 0.96 (read), and with an improved AUPR of 0.99 (instead of 0.92 on the chunk level) and AUROC of 0.99 (instead of 0.90 on the chunk level) (Fig. 3A; Supplementary Table S2). These improvements indicate that the model might be able to use cumulative information per sequencing read to increase overall prediction performance. Additionally, such read-level viability predictions enable inferences from short sequencing reads that had to be excluded from chunk-level analyses; our model achieved good prediction performance on all such previously excluded short reads (in our case, reads shorter than  $<11.5k$  signals;  $n = 166,628$  reads; accuracy: 0.80, AUPR: 0.89, AUROC: 0.87). Given this improved performance, including on previously excluded short sequencing reads, and given that any genomic analysis including taxonomic assignment is usually applied to the unit of the read, we will report all prediction performances in the remaining article on the level of sequencing reads.

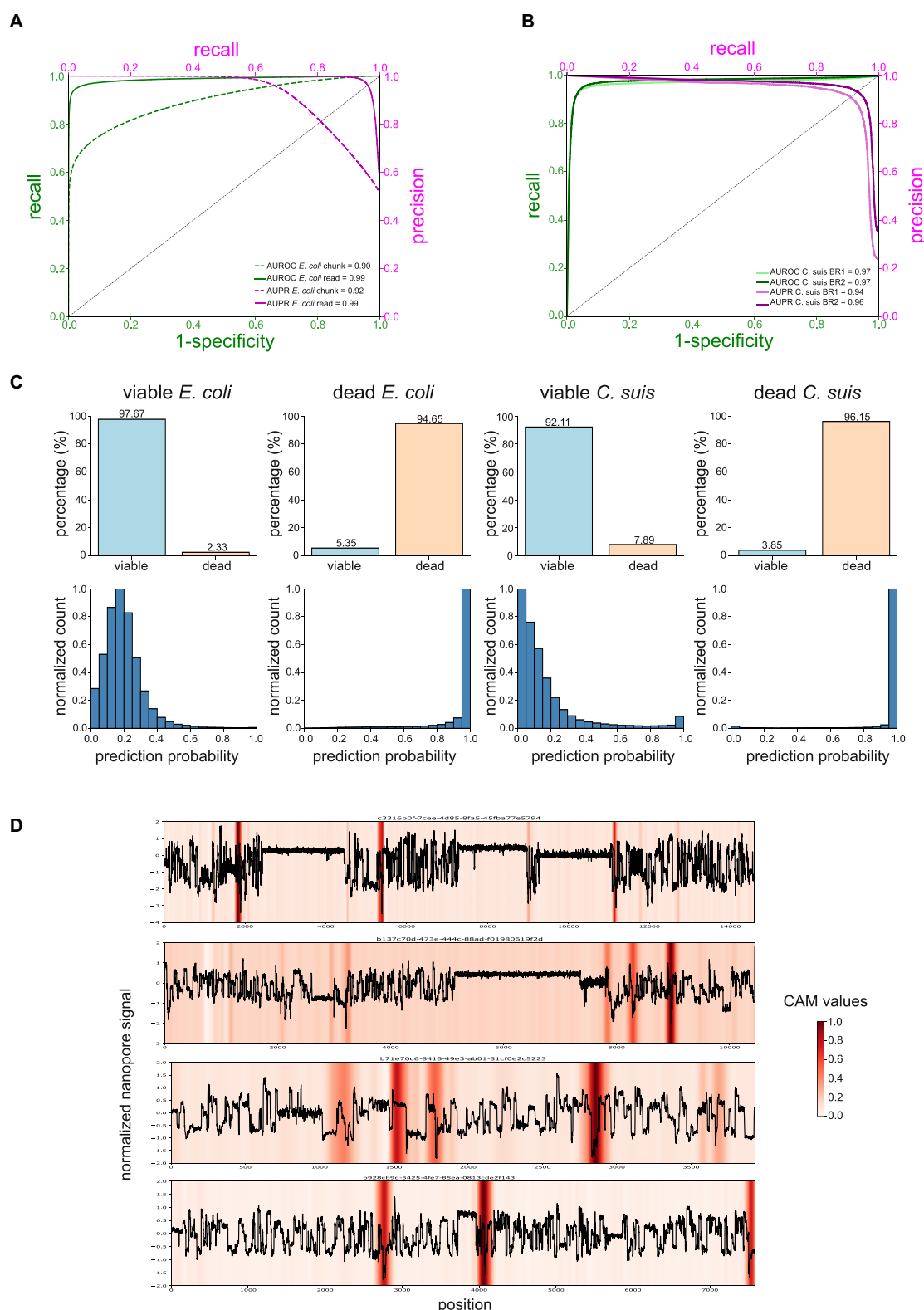
## Application to obligate intracellular *Chlamydia*

We next applied our computational viability framework to distinguish viable from dead *C. suis* cells, an obligate intracellular bacterial species found endemically in the gastrointestinal tract of pigs with high infection rates in pig farms [45, 46]. Like other members of the *Chlamydiaceae* family, these bacteria are defined by a complex biphasic life cycle comprising infectious elementary bodies and dividing reticulate bodies [47]. These unique properties render both cultivation- and vPCR-based approaches for viability estimations complicated [48]. We here took 2 samples from a viable *C. suis* culture as biological replicates (BRs 1 and 2) and subjected them to UV treatment (Materials and Methods). We obtained first insights into their viability using cultivation- and vPCR-based approaches (Materials and Methods). In the case of cultivation, UV-treated *C. suis* cells were unable to form viable inclusions in cell culture, whereas the untreated samples showed high infectivity





**Figure 2:** Explainable AI for interpretability of ResNet1. (A) CAMs leverage the global average pooling (GAP) layer right before the fully connected (FC) layers of the residual neural network to map model interpretability onto the input features; they are generated by aggregating the final convolutional layer's feature maps through a weighted sum, highlighting nanopore signal regions that allow the neural network to make accurate predictions. (B) Exemplary nanopore signal chunks that were classified as “dead” at a prediction probability of  $P > 0.99$  and their CAM values. Higher CAM values indicate stronger feature map activations. (C) Impact of consecutive masking ( $n = 5$  masking events) of the signal region with the highest CAM value per signal chunk (x-axis) on the model's prediction probability (y-axis); 5 different mask sizes (from 100 to 2k signals) were used. (D) Application of a simplified XAI rule that classifies signal chunks according to the presence of a “sudden drop” (Materials and Methods; green: mean signal per chunk; red: threshold for sudden drop definition; yellow: identification of sudden drops in the exemplary signal chunks). (E) Comparison of the performance of the full model (ResNet1) with the simplified XAI rule.



**Figure 3:** Application of the *E. coli*-trained ResNet1 to obligate intracellular *Chlamydia suis*. (A) ResNet1 performance on the *E. coli* test dataset on the chunk level (dashed) and on the sequencing read level (solid) in terms of precision–recall (PR; magenta) and receiver operating characteristic (ROC; green) curves and their respective areas under the curve (AUPR, AUROC). (B) ResNet1 performance on the *C. suis* sequencing reads across 2 biological replicates (BR1 and BR2; light and dark lines, respectively) in terms of PR (magenta) and ROC (green) curves and their respective AUPR and AUROC. (C) Sequencing read-level comparison of ResNet1 classifications of the *E. coli* test and the *C. suis* datasets. Top row: Binary model predictions for viable and dead *E. coli* and *C. suis*, respectively, at the optimized prediction probability threshold of 0.5; bottom row: respective normalized distributions of model prediction probabilities across all sequencing reads. For *E. coli*, all 59,171 “viable” and 72,207 “dead” sequencing reads from the test dataset are shown; for *C. suis*, all 54,853 “viable” and 23,073 “dead” sequencing reads from both biological replicates are shown. (D) Squiggle signal of exemplary *C. suis* sequencing reads that were correctly classified as “dead” and their CAM values; higher CAM values indicate stronger feature map activations (Materials and Methods).

with 8.48e7 (BR1) and 6.23e7 (BR2) inclusion-forming units per milliliter (IFU/mL; [Supplementary Table S3](#)). In the case of vPCR, all samples were processed with and without PMA [48] to quantify the respective amounts of chlamydial DNA with a sensitive *C. suis*-specific quantitative PCR [49, 50] (Materials and Methods). We assessed the difference in copy number per milliliter between PMA-treated and untreated DNA as  $\Delta\log_{10}$ , resulting in 0.82 and 0.81 for the viable samples, as well as 1.51 and 2.02 for the killed samples (BR1 and BR2, respectively; [Supplementary Table S3](#)). These results are comparable to a previous study in which fresh *Chlamydia trachomatis* culture was heat-killed, and an absence of viable *Chlamydia* resulted in a  $\Delta\log_{10}$  *Chlamydia* of 3.01, whereas a viability ratio of 100% resulted in a  $\Delta\log_{10}$  *Chlamydia* of 0.37 [48]. Cultivation and vPCR therefore confirmed that UV treatment had completely inactivated previously viable *C. suis* cells and had strongly reduced the amount of “viable” DNA in both biological replicates. Nanopore sequencing of the viable and dead *C. suis* cells identified 28,210 viable and 8,771 dead (BR1) and 26,643 viable and 14,302 dead (BR2) *Chlamydia*-classified sequencing reads (Materials and Methods; [Supplementary Table S3](#)).

Our *E. coli* data-trained ResNet1 model achieved strong sequencing read-level viability prediction performances across both *C. suis* biological replicates (Materials and Methods). At the previously optimized probability threshold of 0.5, the model achieved an accuracy of 0.93 (both BRs), F1 score of 0.87 and 0.90 (BR1 and BR2), precision of 0.80 and 0.85, recall of 0.95 and 0.96, and specificity of 0.92 and 0.91, respectively ([Supplementary Table S2](#)). Using threshold-independent metrics, the model achieved an AUROC of 0.97 (both BRs) and an AUPR of 0.94 and 0.96, respectively (Fig. 3B). The slightly lower precision obtained by the model's application to BR1 might stem from a slightly higher ratio of viable reads in this biological replicate (76.3% in BR2 vs. 65.1% in BR1). When pooling the model's viability predictions across the 2 biological replicates, the percentage of correctly classified sequencing reads (Fig. 3C, top row) and the prediction probability distributions across reads (Fig. 3C, bottom row) are comparable with sequencing read-level performances in the original *E. coli* test dataset (Fig. 3C, left). This shows a certain degree of generalizability of our viability model beyond taxonomic boundaries despite the model being trained only on *E. coli* data. While both bacterial species are Gram-negative, they substantially differ in their ecology and life cycles, suggesting a potential applicability of deep models to nanopore squiggle data for taxonomy-agnostic viability predictions. The application of XAI to the *C. suis* sequencing reads further shows that the CAMs also highlight sudden drops in the *C. suis* squiggle data (Fig. 3D), suggesting that UV exposure led to similar damage in both bacterial species. The killing method or, more generally, the source of degradation might therefore be the main determinant of viability-predictive features in nanopore squiggle data.

## Viability inference after antibiotic exposure of a mock community

To test whether our ResNet1 model that was trained on UV-killed *E. coli* could be used for viability predictions using a different killing method, we generated 2 biological replicates (BR1 and BR2) of a mock community of carbenicillin-susceptible *E. coli* and carbenicillin-resistant *Klebsiella oxytoca*. Nanopore sequencing of the mock community after carbenicillin exposure resulted in 239,316 and 240,178 *E. coli* sequencing reads (number of reads mapping to the *Escherichia* genus in BR1 and BR2) and in 991,019 and 954,602 *K. oxytoca* sequencing reads for the experiment (number of reads mapping to the *Klebsiella* genus in BR1 and BR2; Mate-

rials and Methods). Across both species and biological replicates, most of the sequencing reads were classified as viable (*E. coli*: 82.2% and 76.4%; *K. oxytoca*: 80.8% and 87.4%, for BR1 and BR2, respectively), showing that ResNet1 could not distinguish susceptible from resistant bacteria after antibiotic exposure. This finding supports our hypothesis that the killing method or, more generally, the source of degradation might be the main determinant of viability-predictive features in nanopore squiggle data.

We therefore explored if we could train a new model using the same previously optimized ResNet1 architecture to accurately predict viability after antibiotic exposure. We first generated a clean antibiotic exposure training dataset of a single bacterial species by nanopore-sequencing the susceptible *E. coli* strain before and after exposure; similar to the UV-treated *E. coli*, we additionally made sure to (i) only sequence cell-free DNA for the dead viability class by stringent centrifugation and filtering of the supernatant and to (ii) mainly sequence DNA from intact cells for the viable class by only processing the pellet after centrifugation (Materials and Methods). The newly trained antibiotic ResNet1 model achieved a test dataset accuracy of 0.73 on the sequencing read level, with an AUPR of 0.98 and an AUROC of 0.87, and a performance on a held-out biological replicate of 0.68 accuracy, with an AUPR of 0.95 and an AUROC of 0.80 (based on 16,810 viable and 103,120 dead sequencing reads that were classified as *Escherichia*; [Supplementary Table S4](#); Materials and Methods).

We then applied this antibiotic ResNet1 model to the antibiotic exposure mock community; this time, most of sequencing reads from the susceptible *E. coli* were classified as dead (75.7% and 75.9%, for BR1 and BR2), while the majority of *K. oxytoca* was still correctly classified as viable (70.1% and 71.7%). The application of our XAI framework showed that this new antibiotic ResNet1 did not detect any sudden drops in the squiggle data that were previously identified by the UV ResNet1's CAMs in dead sequencing reads ([Supplementary Fig. S4](#) for a few examples), showing that this new model seemed to have identified different squiggle signatures indicative of degradation through antibiotic exposure. These preliminary results suggest that our viability model can be tuned to classify viability in different degradation contexts and that such models can potentially be applied to separate resistant from susceptible bacteria in mixed microbial communities.

## AI- and nanopore-empowered viability-resolved metagenomics

While metagenomic approaches provide the unique opportunity of generating *de novo* assemblies and potentially complete microbial genomes to explore the “microbial dark matter” as well as to infer potential functions such as metabolic and virulence potential [8, 9], they have suffered from their inability to differentiate between viable and dead microorganisms [2, 15]. Such viability inferences can, however, distort any microbial inference, ranging from assessing ecosystem functions of environmental microbiomes to inferring the virulence of potential pathogens. As established viability-resolved metagenomic approaches are labor-intensive as well as biased and lack sensitivity (e.g., [23]), we here show first evidence that a fully computational framework based on residual neural networks with convolutional data processing layers can leverage raw nanopore signal data, also known as squiggle data, to make accurate inferences about microbial viability. Using experimentally killed bacterial cultures and a simple mock microbial community, we show that such models can infer viability from sequencing reads at high accuracy, potentially

allowing for simultaneous taxonomic and viability classifications in metagenomic datasets.

We first leverage microbial degradation through UV exposure to show that such viability models can make accurate predictions in a taxonomy-agnostic manner; a model that has only ever seen *E. coli* squiggle data can predict viability in UV-exposed *Chlamydia* at high recall and specificity ( $>0.9$ ). The application to estimate the viability of pathogenic *Chlamydia* is hereby of potentially immediate interest to veterinary scientists since traditional methods for assessing the pathogen's viability have been labor-intensive and suffered from inherently high false-negative rates; more research is, however, needed to assess to what extent this model can capture natural degradation in *Chlamydia*.

Our subsequent XAI analyses point to the potential role of UV-induced DNA backbone damage for achieving accurate model predictions in both species; the simplified XAI rule is, however, not sufficient to correctly classify the majority of “dead” signal chunks (recall of 0.39), which means that the residual neural network has apparently identified additional more subtle signal patterns that allow the full model to make more sensitive predictions (recall of 0.74) while maintaining specificity and precision. More research is therefore needed to fully understand the biological, physical, or chemical underpinnings of our viability model predictions; we, however, anticipate that future experiments for squiggle data generation, including on different taxonomic groups and killing methods, will help us tease apart the origins of our current XAI results.

Besides exploring the underlying rules of squiggle signal patterns, the generalizability of such squiggle-based viability models also needs to be assessed in more detail, including for a wider breadth of microbial taxa such as spore-forming bacteria or fungi, and for a variety of degradation sources and intensities. Our preliminary results indicate that our deep model that can accurately predict UV-induced degradation cannot distinguish susceptible and resistant bacteria in an antibiotic exposure experiment of a simple mock community. We, however, show that training the same residual neural network architecture on newly generated killing method-specific squiggle data can achieve good accuracy in the respective mock community ( $>0.7$ ). While this application shows the limits of our current models' generalizability, we also argue that clinical metagenomics might already profit from such antibiotic-specific viability inferences: as previously discussed, certain disinfection methods or systemic antibiotics in the clinical setting often kill the bacteria before the DNA is destroyed [15, 17], leading to potential false-positive pathogen detections using metagenomics; our viability model could give additional information on the antibiotic exposure's impact, and confidence could be increased by accumulating viability evidence across sequencing reads per pathogen of interest. In addition, we hypothesize that models could be trained to detect damage induced by suboptimal antibiotic usage, which has been implicated in the emergence of new antibiotic resistances [51].

In order to achieve true metagenomic applications in the future, the training data of a single model could be diversified in terms of taxonomy and degradation, potentially increasing the generalizability of its viability inferences. We also anticipate that quantitative AI modeling has the potential to inform more differentiated viability assessments, which might help quantify or time degradation events and even decipher the relevance of dormancy and metabolic inactivity in metagenomic studies [52, 53]. We envision many potential benefits of such a widely applicable computational framework for microbial viability inference, including for applications in environmental, veterinary, and clinical

settings. As is the case for epigenetic inferences [36–38], the viability inference-enabling squiggle data are a complementary output of any nanopore sequencing experiment of native DNA (and RNA), which is then usually basecalled and archived for future re-basecalling after potential basecalling model improvements. This means that any future nanopore-based metagenomic study could make viability predictions for free without additional costs and laboratory work, and any existing archived nanopore data could be assessed in terms of its microorganisms' viability—which would allow us to quantify the impact of dead microorganisms on metagenomic analyses in a diversity of datasets and ecological settings, as well as further explore factors such as species and environment specificity.

## Materials and Methods

### Viability inference explorations in UV-exposed *E. coli*

#### Training data generation

We cultured *E. coli* K12 in 200 mL Luria–Bertani (LB) medium for 24 hours at 37°C to reach the log phase of the growth curve. The culture was then used to inoculate four 200 mL LB media in 1-L Erlenmeyer flasks, which were again incubated for 24 hours to reach the growth log phase. One of the media was used as viable control, that is, DNA was extracted from 750  $\mu$ L of the living culture using the spin column-based QIAGEN PowerSoil Pro Kit, following the manufacturers' instructions. The remaining 3 cultures were killed by one of the following stressors: UV irradiation at 254 nm for 15 minutes, heat shock at 120°C for 5 minutes, or bead beating for 30 minutes. To then separate extracellular DNA from cell debris and intact bacterial cells, we centrifuged the media for 10 minutes at 4,000  $\times$  g and filtered the supernatant through 0.2- $\mu$ m filters. The resulting extracellular DNA was subsequently kept at room temperature for 5 days to simulate the natural accumulation of DNA degradation. DNA from dead bacteria was extracted from these samples using the same extraction approach following the QIAGEN PowerSoil Pro Kit protocol, but the first lysis buffer step was omitted since cell lysis had already happened.

We then used Oxford Nanopore Technologies' Rapid Barcoding library preparation kit (RBK114-24 V14), R10.4.1 MinION flow cells, and MinKNOW software v23.04.5 for shotgun nanopore sequencing of the “viable” and “dead” DNA. We used 4 barcodes for each sample, resulting in DNA input of 800 ng and 218 ng for the preparation of the “viable” and “dead” library, respectively. We ran each library for 24 hours, using 2 different flow cells to avoid any cross-contamination, and filtered the resulting nanopore data at a minimum read length of 20 b. Raw nanopore data were created using the standard translocation speed of 400 b/s and a sampling frequency of 5 kHz. We applied Dorado [54] SUP-basecalling model v4.2.0 ([dna\\_r10.4.1\\_e8.2\\_400bps\\_sup@v4.2.0](#)) and 6mA-aware SUP-basecalling (6mA@v1) to all nanopore reads that had passed internal data quality thresholds to obtain *E. coli* DNA sequence data. We subsequently removed sequencing adapters and barcodes using Porechop v0.2.3 [55].

#### Model training

We tested the implementation of different residual neural networks and transformer architectures to predict the binary viability state from the raw nanopore data (0 = viable; 1 = dead). The first residual neural network, ResNet1, consists of 4 layers, each containing 2 bottleneck blocks. Each bottleneck block consists of convolutional layers, batch normalization, and a rectified



linear unit (ReLU) activation function. Each of the 4 layers consists of an increasing number of convolutional channels: 20, 30, 45, and 65, respectively, followed by global average pooling and a fully connected layer, resulting in 66,916 parameters. The model then uses a softmax function to convert logits, the raw outputs from the fully connected layer, into predicted probabilities ranging from 0 to 1. We evaluated the training of the model using the Adam optimizer for mini-batch gradient descent at 3 different LRs (1e-3, 1e-4, and 1e-5), training the model up to 1,000 epochs and at a batch size of 1,000 signal chunks. We initialized the model using Kaiming initialization. For ResNet2, we increased the number of convolutional channels to 40, 60, 90, and 135, respectively, resulting in 1,828,777 parameters. For ResNet3, we increased the number of convolutional channels to 512, 30, 45, and 67, respectively, resulting in 2,479,140 parameters. The transformer model was based on a positional encoding, a convolutional layer with a channel number of 24, and 1 block of 1 attention head, resulting in 219,586 parameters.

We processed the *E. coli* squiggle data by excluding the first 1,500 signal points (potential noise, adapter sequences, or barcodes), then cutting them into signal chunks of 10k signals, and separated the chunks into balanced training (60%), validation (20%), and test (20%) sets along each original sequencing read to avoid that signal chunks from the same read would end up in the same dataset. We pooled the viable and dead signal chunks to obtain exactly balanced training, validation, and test sets. For normalizing each chunk, we subtracted the median per chunk and divided it by the median absolute deviation (MAD) to make the signal data robust to outliers. We then scaled the signal by the MAD scaling factor 1.4826 and replaced outliers exceeding 3.5 times the scaled MAD by the mean of their 2 neighboring values.

We also trained ResNet1 on the basecalled nanopore data (with or without 6 mA basecalling) of viable and dead *E. coli* at a standardized chunk size of 800 b, which roughly corresponds to a signal chunk size of 10k signals. For encoding, we used a one-hot encoding method to turn DNA sequences into unique binary vectors. We then concatenated and saved these encoded sequences as tensors for training and testing. We finally trained ResNet1 on signal chunks of different signal lengths, ranging from 1k to 20k signals.

For logistic regression training, we used the LogisticRegression class from scikit-learn v1.2.2 with default parameters.

## Explainable AI

We utilized CAMs to identify and visualize signal regions that influenced the model's decision-making. As feature maps from the final convolutional layer undergo a global average pooling layer where each map is averaged and concatenated, we can calculate the weighted sum of these feature maps using the weights of the fully connected layer and project it back onto the preprocessed signal [42]. To do so, we implemented CAM in Python/PyTorch. During the forward pass, we ensure that the feature maps from the last convolutional layer are captured. To compute the CAM, we use the weights of the model's output layer for the class of interest, multiplying these weights with the corresponding feature maps and then summing them up. We convert the resulting CAM to an array and normalize its values to a range of 0 to 1. We then overlay the CAM on the original input signal to identify the regions most influential in the model's decision-making process. We additionally used the Remora API to match raw nanopore data to the corresponding Dorado-basecalled bases, to then manually investigate any obvious sequence abnormalities in the CAM regions.

For consecutive masking of the CAM regions with the highest CAM values, we used Python to first obtain and normalize the CAM values of all true-positive signal chunks at  $P > 0.5$  ( $n = 286,179$ ), identify the maximum value, mask the signal region (i.e., setting to zero after MAD normalization) at a specified mask size (between 100 and 2k signals) centered on the maximum-CAM signal, and obtain updated prediction probabilities. We repeated this masking step 5 times and calculated the confidence interval at each masking step: We obtained the mean and standard error of the mean (SEM) of the newly calculated prediction probabilities across all signal chunks to calculate the 95% confidence interval at  $\text{mean} \pm 1.96 * \text{SEM}$ . We plotted the results using matplotlib.

We next used Python to develop an algorithm to obtain a simplified XAI rule to distinguish “dead” from “viable” signals chunks based on our CAM results by identifying the presence of at least 1 sudden drop in the chunk. To identify those sudden drops, we first calculated the mean and standard deviation (SD) of each signal chunk and found that a scaling factor of 3 identified most manually detected sudden drops at a vertical threshold of  $\text{mean} - 3 * \text{SD}$ .

## Sequencing read-level inferences

To evaluate read-level performance, we leveraged the inherent flexibility of ResNet architectures to handle variable-length input. Prior to all read-level analyses, we removed chimeric reads using information from Dorado basecalling. Dorado splits chimeric reads and tags the resulting child reads with a “pi” tag (parent\_read\_id) in the BAM file, indicating their origin from an unsplit read. Using the pysam Python package, we extracted the read IDs of the reads carrying the “pi” tag to exclude chimeric reads from our downstream inference analysis. The nonchimeric reads were subsequently filtered according to Kraken2 v2.1.4-based taxonomic assignment to the *Escherichia* genus [56]. All sequencing read-level inferences were made only on nonchimeric, correctly taxonomically classified (on the genus level) reads.

## Application to obligate intracellular *Chlamydia*

The *C. suis* strain S45 (kindly provided by J. Storz) was cultured as described by Leonard et al. [57]. Briefly, chlamydiae were grown in the epithelial rhesus monkey kidney cell line LLM-MK2 (provided by IZSLER) and prepared as semi-purified stock by scraping infected cells into the supernatant and removing cellular debris by centrifugation at 500 *g* for 10 min. Chlamydiae were then pelleted (10,000 × *g*, 45 minutes) and resuspended in sucrose phosphate glutamate (SPG) buffer. To determine the viability of this stock, 1 aliquot was thawed on ice and separated into 2 tubes, of which 1 was UV-inactivated using a Hoefer UVC 500 Ultraviolet Crosslinker: briefly, samples were kept on ice and exposed to 8 watts of UV light at 12.5 cm for 30 minutes, similar to previously described protocols [58]. All samples were prepared in biological duplicates. UV-inactivated chlamydiae were then further incubated at room temperature for 48 hours prior to further processing, while viable chlamydiae were immediately processed. The resulting 4 samples were divided into subsamples and used for cultivation, vPCR, and nanopore sequencing.

For viability determination by culture, subsamples (approximately 1e7 IFU) were used to infect 4 glass coverslips (13 mm in diameter; ThermoScientific) in 24-well plates (TPP Techno Plastic Product AG) seeded to confluence with LLC-MK2 cells [59]. For the “viable” subsamples, a 1:1,000 dilution was performed prior to inoculation. The infection was then enhanced by centrifugation for 1 hour at 25°C (1,000 × *g*). After 48 hours of incubation

at 37°C (5% CO<sub>2</sub>), cultures were fixed for 10 minutes in ice-cold methanol. If cultures were free of chlamydiae, 1 well was scraped and transferred to fresh monolayers up to 3 times to confirm negativity as described [60]. Coverslips were then processed using a well-established immunofluorescence assay [61]. Briefly, DNA was stained with 1 µg/mL 4',6-diamidino-2'-phenylindole dihydrochloride (DAPI; Molecular Probes). In parallel, inclusions were labeled with a *Chlamydiaceae*-specific primary antibody (*Chlamydiaceae* LPS, Clone ACI-P; Progen), which was diluted 1:200 in blocking solution consisting of 1% bovine serum albumin (BSA) in phosphate-buffered saline (PBS; GIBCO, Invitrogen). Inclusions were visualized with Alexa Fluor 488 goat anti-mouse (Molecular Probes) diluted 1:500 in blocking solution. As a final step, coverslips were washed with PBS, mounted with FluoreGuard (Hard Set; ScyTek Laboratories) on glass slides, and inclusions determined using a Leica DMLB fluorescence microscope (Leica Microsystems) and a 10× ocular objective (Leica L-Plan 10×/25 M; Leica Microsystems). In parallel, a 3-fold dilution series of the sample was performed in 96-well plates (TPP) and processed as above. The number of IFU/mL for the whole samples was then calculated based on the number of inclusions detected using the Nikon Ti Eclipse epifluorescence microscope at a 20× magnification [59].

For vPCR, subsamples (approximately 2e7 IFU) were taken and mixed with 200 µL SPG and 100 µL PMA enhancer for Gram-Negative Bacteria (5×; Biotium). PMAxx (Biotium) at a final concentration of 50 µM was added ("PMA-treated") or not ("untreated") to the subsamples. Samples were then exposed to a 650-W light source using a PMA-Lite LED Photolysis Device (Biotium) for 5 minutes, followed by 2 minutes on ice and additional light exposure for 5 minutes [48]. For vPCR as well as nanopore sequencing (subsamples with approximately 5e8 IFU), DNA was extracted using the DNeasy Blood and Tissue Kit (QIAGEN) according to the manufacturer's instructions. The amount of chlamydial DNA was quantified with a sensitive *C. suis* quantitative PCR [62], based on a standard curve with recombinant plasmid containing the amplicon target and calculated for the whole sample. For subsequent nanopore sequencing of the DNA extracts (viable: 13–16 ng/µL; dead: 1.9–2.8 ng/µL), we followed the same approach as described for *E. coli* above. Each viable sample was sequenced using 1 barcode (input volume of 10 µL) and each dead sample using 3 barcodes (input volume of 30 µL). Sequencing reads were processed following the same steps used for the *E. coli* dataset. Chimeric reads were identified using the "pi" tag added by Dorado and removed with the pysam Python package. Adapter and barcode trimming was performed using Porechop, and taxonomic classification was conducted with Kraken2 v2.1.4, retaining only reads classified as *Chlamydia*.

## Antibiotic exposure experiment

To generate data from a simple mock community, a carbenicillin-susceptible *E. coli* strain and an ESBL-producing *K. oxytoca* strain (ATCC 700324) were cultured on Müller–Hinton (MH) agar at 37°C for 20 hours in 2 biological replicates. Susceptibility to penicillin as an approximation of carbenicillin resistance was confirmed by VITEK2 and growth curve analysis at 100 and 200 ng/µL. Species identity was confirmed by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS) and Multi-Locus Sequence Typing (MLST). Cultures were inoculated at OD<sub>600</sub> = 0.01 in 10 mL MH broth and incubated for 5 hours. Carbenicillin (200 ng/µL) was added during logarithmic growth, followed by 20 hours of further incubation. DNA was ex-

tracted directly from 250 µL of uncentrifuged culture using the same DNA extraction approach as described for UV-exposed *E. coli*.

To generate clean training data for an antibiotic exposure-specific model, 2 biological replicates of the same *E. coli* strain were used. For the viable samples, DNA was extracted from centrifuged pellets (4,000 × g, 10 minutes) before antibiotic treatment, respectively. For the dead samples, DNA was extracted from the filtered supernatant (0.2 µm) 20 hours after carbenicillin exposure using the same DNA extraction approach as described for UV-exposed *E. coli*.

Nanopore sequencing and data processing were done as described for *E. coli* above. All inferences were done on the sequencing read level and restricted to nonchimeric reads classified as *Escherichia* or *Klebsiella*, respectively. Model training was also done as described above, using the optimized ResNet1 architecture, but chunk size was set to 5k instead of 10k signals to increase the size of the training dataset. The final dataset consisted of 330,000 chunks for training, 110,000 for validation, and 110,000 for testing. The model was trained for up to 600 epochs using the Adam optimizer with a learning rate of 1e-5. The best validation accuracy was achieved at epoch 550.

## Availability of Supporting Source Code and Requirements

Project name: Squiggle4Viability

Project homepage: <https://github.com/Genomics4OneHealth/Squiggle4Viability.git>

Operating system(s): Platform independent

Programming language: Python

License: MIT

An archival version of the code is available via Software Heritage [63].

## Additional Files

**Supplementary Table S1.** UV-killed *E. coli* viability inferences of deep neural network and logistic regression models. Test dataset performance metrics of residual neural networks ("ResNet"), transformer architectures, and logistic regression models, trained on various data modalities (nanopore squiggle "Signal" or base-called DNA "Nucleotide" sequence) at various signal chunk sizes ("Length") and on sequencing read length or translocation speed ("Trans speed").

**Supplementary Table S2.** Sequencing read-level viability inferences of UV ResNet1. Performance metrics across sequencing reads of the ResNet1 model trained on UV-killed *E. coli* for the *E. coli* test dataset and 2 biological replicates (BR1 and BR2) of UV-killed and viable *Chlamydia suis*. The number of total reads is the number of sequencing reads after processing of the nanopore sequencing data by Porechop; the number of genus-classified reads is the number of sequencing reads that map to the *Escherichia* or *Chlamydia* genus, respectively, using Kraken2 (Materials and Methods).

**Supplementary Table S3.** Cultivation, viability PCR (vPCR), and nanopore sequencing metrics of UV-killed *Chlamydia suis*. Cultivation titer in number of inclusion forming units (IFUs) per milliliter, PMA-untreated vPCR reflecting total *Chlamydia* content, PMA-treated vPCR reflecting viable *Chlamydia* content and Δlog<sub>10</sub> of the PMA-treated copy number divided by the PMA-untreated copy number reflecting overall viability, total number of nanopore sequencing reads after Porechop processing, and

number of *Chlamydia*-classified reads using Kraken2 (Materials and Methods), of 2 biological replicates (BR1 and BR2) of viable and dead *C. suis*.

**Supplementary Table S4.** Sequencing read-level viability inferences of antibiotic exposure ResNet1. Performance metrics across sequencing reads of the ResNet1 model trained on antibiotic-exposed *E. coli* for the *E. coli* test dataset and held-out biological replicates (BRs). The number of *E. coli* reads is the number of sequencing reads after processing of the nanopore sequencing data by Porechop and mapping to the *Escherichia* genus using Kraken2 (Materials and Methods).

**Supplementary Fig. S1.** Training and validation loss across deep neural network architectures tested for nanopore squiggle signal-based viability inference. (A–C) Model loss of ResNet1 at learning rates (LRs) of 1e-3, 1e-4, and 1e-5. (D–F) Model loss of ResNet2 at LRs of 1e-3, 1e-4, and 1e-5. (G–I) Model loss of ResNet3 at LRs of 1e-3, 1e-4, and 1e-5. (J–L) Model loss of the transformer models at LRs of 1e-3, 1e-4, and 1e-5. The solid blue line indicates the training loss, the solid red line indicates the validation loss, and the dashed line indicates the minimum validation loss from the final ResNet1, LR = 1e-4, model.

**Supplementary Fig. S2.** Training and validation loss of ResNet1 at various signal chunk sizes. The signal chunk size varies: (A) 1k, (B) 5k, (C) 7k, (D) 10k, (E) 12k, and (F) 20k. The solid blue line indicates the training loss, the solid red line indicates the validation loss, and the dashed line indicates the minimum validation loss from the final ResNet1 model using a signal chunk size of 10k.

**Supplementary Fig. S3.** Exemplary drops in ResNet1 prediction probabilities in nanopore signal chunks after consecutive masking of the signal region with the respectively highest CAM value. *Figure headers:* signal chunk ID/total number of masked signal values/prediction probability per signal chunk “prob.” *Left to right:* Five exemplary nanopore signal chunks (length of 10k signals). *Top to bottom:* Consecutive masking of 200 signal values per masking event (Materials and Methods). *Legends:* Red-colored CAM value visualizations; higher CAM values indicate stronger feature map activations.

**Supplementary Fig. S4.** Exemplary nanopore signal patterns of antibiotic-killed *E. coli* sequencing reads and XAI Class Activation Mapping (CAM). *Legend:* Red-colored CAM value visualizations; higher CAM values indicate stronger feature map activations.

## Abbreviations

AUC: area under the curve; AUPR: area under the precision–recall curve; AUROC: area under the receiver operating characteristic curve; BSA: bovine serum albumin; CAM: Class Activation Map; EMA: ethidium monoazide; FC: fully connected; GAP: global average pooling; LB: Luria–Bertani; LR: learning rate; MAD: median absolute deviation; MAG: metagenome-assembled genome; MH: Müller–Hinton; PBS: phosphate-buffered saline; PMA: propidium monoazide; ResNet: Residual Neural Network; ROC: receiver operating characteristic; SEM: standard error of the mean; SPG: sucrose phosphate glutamate; vPCR: viability PCR; XAI: explainable artificial intelligence.

## Acknowledgments

The authors thank the laboratory of the Research Unit of Comparative Microbiome Analysis at Helmholtz Munich, Germany, especially Cornelia Galonska, for their support in processing the *Escherichia coli* samples; the laboratory of the Institute of Veterinary Pathology, Switzerland, especially Theresa Pesch, for their support

in processing the *Chlamydia suis* samples; and Valentin Rauscher for his help in training several deep models during his internship in the Urban research group at Helmholtz Munich.

## Author Contributions

Harika Ürel (Formal analysis, Software, Validation, Writing—original draft preparation, Writing—review & editing, Visualization), Sabrina Benassou (Formal analysis, Visualization, Writing—review & editing), Hanna Marti (Investigation, Resources, Writing—original draft preparation, Writing—review & editing), Tim Reska (Investigation, Writing—review & editing), Ela Sauerborn (Investigation, Writing—review & editing), Yuri Pinheiro Alves De Souza (Investigation), Albert Perlas (Investigation), Enrique Rayo (Investigation), Michael Biggel (Investigation), Stefan Kesselheim (Supervision, Writing—review & editing), Nicole Borel (Supervision), Edward J. Martin (Formal analysis), Constanza B. Venegas (Investigation), Michael Schlöter (Supervision, Resources, Writing—review & editing), Kathrin Schröder (Investigation), Jana Mittelstrass (Investigation), Simone Prospero (Investigation), James M. Ferguson (Formal analysis), and Lara Urban (Conceptualization, Supervision, Formal analysis, Resources, Funding acquisition, Writing—original draft preparation, Writing—review & editing)

## Funding

This study was funded by a Helmholtz Principal Investigator Grant awarded to L.U. by Helmholtz AI. The *Chlamydia suis* research was financed by the Vontobel Foundation (grant number 1336/2004; Zürich, Switzerland). Computational resources were provided by Helmholtz Munich and by the Helmholtz Association Initiative and Networking Fund HAICORE partition at the Forschungszentrum Jülich. H.U. was supported by the Helmholtz Association under the joint research school “HIDSS-006–Munich School for Data Science@Helmholtz, TUM&LMU.” E.M. was supported by an EAST-BIO studentship, funded by BBSRC Grant Number BB/M010996/1, and an STFC Food Network+ Scoping Grant. S.B. and S.K. were supported by Helmholtz AI’s Helmholtz Association Initiative and Networking Fund. L.U. was supported by funding from the University of Zurich. L.U., H.U., E.S., A.P., and J.M.F. have previously received travel and accommodation funding to speak at Oxford Nanopore Technologies’ conferences.

## Data Availability

All nanopore sequencing raw data in pod5 file format are available via the GigaScience database, GigaDB [64], and the SquidBase database (accession number SQB000009) [65] and have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB76420. The checkpoint (ckpt) files of all machine learning models are available at [66]. DOME-ML (Data, Optimization, Model and Evaluation in Machine Learning) annotations are available in the DOME registry via accession 66r9zbj2zc [67].

## Competing Interests

The authors declare that they have no competing interests.



## References

- Lewis WH, Tahon G, Geesink P, et al. Ettema TJG. Innovations to culturing the uncultured microbial majority. *Nat Rev Micro*. 2021;19(4):225–40. <https://doi.org/10.1038/s41579-020-00458-8>.
- Lloyd KG, Steen AD, Ladau J, et al. Phylogenetically novel uncultured microbial cells dominate Earth microbiomes. *mSystems*. 2018;3(5). <https://doi.org/10.1128/msystems.00055-18>.
- Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, et al. A catalog of reference genomes from the human microbiome. *Science*. 2010;328(5981):994–99. <https://doi.org/10.1126/science.1183605>.
- Sauerborn E, Corredor NC, Reska T, et al. Detection of hidden antibiotic resistance through real-time genomics. *Nat Commun*. 2024;15:5494. <https://doi.org/10.1038/s41467-024-49851-4>.
- Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*. 2004;68(4):669–85. <https://doi.org/10.1128/MMBR.68.4.669-685.2004>.
- Pace NR. Mapping the tree of life: progress and prospects. *Microbiol Mol Biol Rev*. 2009;73(4):565–76. <https://doi.org/10.1128/MMBR.00033-09>.
- Brooks JP, Edwards DJ, Harwich MD, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol*. 2015;15:66. <https://doi.org/10.1186/s12866-015-0351-6>.
- Dick GJ, Andersson AF, Baker BJ, et al. Community-wide analysis of microbial genome sequence signatures. *Genome Biol*. 2009;10(8):R85. <https://doi.org/10.1186/gb-2009-10-8-r85>.
- Quince C, Walker AW, Simpson JT, et al. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol*. 2017;35(9):833–44. <https://doi.org/10.1038/nbt.3935>.
- Sereika M, Kirkegaard RH, Karst SM, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nat Methods*. 2022;19(7):823–26. <https://doi.org/10.1038/s41592-022-01539-7>.
- Liu L, Yang Y, Deng Y, et al. Nanopore long-read-only metagenomics enables complete and high-quality genome reconstruction from mock and complex metagenomes. *Microbiome*. 2022;10(1):209. <https://doi.org/10.1186/s40168-022-01415-8>.
- Jain M, Olsen HE, Paten B, et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol*. 2016;17:239. <https://doi.org/10.1186/s13059-016-1103-0>.
- Pagès-Gallego M, de Ridder J. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. *Genome Biol*. 2023;24(1):71. <https://doi.org/10.1186/s13059-023-02903-2>.
- Urban L, Perlas A, Francino O, et al. Real-time genomics for One Health. *Mol Syst Biol*. 2023;19(8). <https://doi.org/10.15252/msb.202311686>.
- Nogva HK, Drømtorp SM, Nissen H, et al. Ethidium monoazide for DNA-based differentiation of viable and dead bacteria by 5'-nuclease PCR. *BioTechniques*. 2003;34(4):804–8, 810, 812–13. <https://doi.org/10.2144/03344rr02>.
- Carini P, Marsden P, Leff J, et al. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nat Microbiol*. 2017;2:16242. <https://doi.org/10.1038/nmicrobiol.2016.242>.
- Hellmann KT, Tuura CE, Fish J, et al. Viability-resolved metagenomics reveals antagonistic colonization dynamics of *Staphylococcus epidermidis* strains on preterm infant skin. *mSphere*. 2021;6(5). <https://doi.org/10.1128/mSphere.00538-21>.
- Reska T, Pozdniakova S, Borràs S, et al. Air monitoring by nanopore sequencing. *ISME Commun*. 2024;4(1):ycae099. <https://doi.org/10.1093/ismeco/ycae099>.
- Morré SA, Sillekens PT, Jacobs MV, et al. Monitoring of *Chlamydia trachomatis* infections after antibiotic treatment using RNA detection by nucleic acid sequence based amplification. *Mol Pathol*. 1998;51(3):149–54. <https://doi.org/10.1136/mp.51.3.149>.
- Herman L. Detection of viable and dead *Listeria monocytogenes* by PCR. *Food Microbiol*. 1997;14(2):103–10. <https://doi.org/10.1006/fmic.1996.0077>.
- Min J, Baeumner AJ. Highly sensitive and specific detection of viable *Escherichia coli* in drinking water. *Anal Biochem*. 2002;303(2):186–93. <https://doi.org/10.1006/abio.2002.5593>.
- Urban L, Holzer A, Baronas JJ, et al. Freshwater monitoring by nanopore sequencing. *eLife*. 2021;10. <https://doi.org/10.7554/eLife.61504>.
- Kumar SS, Ghosh AR. Assessment of bacterial viability: a comprehensive review on recent advances and challenges. *Microbiology (Reading)*. 2019;165(6):593–610. <https://doi.org/10.1099/mic.0.000786>.
- Taylor-Brown A, Madden D, Polkinghorne A. Culture-independent approaches to chlamydial genomics. *Microb Genom*. 2018;4(2):e000145. <https://doi.org/10.1099/mgen.0.000145>.
- Nocker A, Camper AK. Novel approaches toward preferential detection of viable cells using nucleic acid amplification techniques. *FEMS Microbiol Lett*. 2009;291(2):137–42. <https://doi.org/10.1111/j.1574-6968.2008.01429.x>.
- Gomez-Silvan C, Leung MHY, Grue KA, et al. A comparison of methods used to unveil the genetic and metabolic pool in the built environment. *Microbiome*. 2018;6:71. <https://doi.org/10.1186/s40168-018-0453-0>.
- Sheridan GE, Masters CI, Shallcross JA, et al. Detection of mRNA by reverse transcription-PCR as an indicator of viability in *Escherichia coli* cells. *Appl Environ Microb*. 1998;64(4):1313–18. <https://doi.org/10.1128/AEM.64.4.1313-1318.1998>.
- Wong VY, Duval MX. Inter-laboratory variability in array-based RNA quantification methods. *Genomics Insights*. 2013;6:13–24. <https://doi.org/10.4137/GEI.S11909>.
- Janssen KJH, Dirks JAMC, Dukers-Muijters NHTM, et al. Review of *Chlamydia trachomatis* viability methods: assessing the clinical diagnostic impact of NAAT positive results. *Expert Rev Mol Diagn*. 2018;18(8):739–47. <https://doi.org/10.1080/14737159.2018.1498785>.
- Nocker A, Cheung CY, Camper AK. Comparison of propidium monoazide with ethidium monoazide for differentiation of live vs. dead bacteria by selective removal of DNA from dead cells. *J Microbiol Methods*. 2006;67(2):310–20. <https://doi.org/10.1016/j.mimet.2006.04.015>.
- Emerson JB, Adams RI, Román CMB, et al. Schrödinger's microbes: tools for distinguishing the living from the dead in microbial ecosystems. *Microbiome*. 2017;5(1):86. <https://doi.org/10.1186/s40168-017-0285-3>.
- Bae S, Wuertz S. Discrimination of viable and dead fecal Bacteroidales bacteria by quantitative PCR with propidium monoazide. *Appl Environ Microb*. 2009;75(9):2940–44. <https://doi.org/10.1128/AEM.01333-08>.
- Courcelle J, Donaldson JR, Chow KH, et al. DNA damage-induced replication fork regression and processing in *Escherichia coli*. *Science*. 2003;299(5609):1064–67. <https://doi.org/10.1126/science.1081328>.
- Setlow RB, Swenson PA, Carrier WL. Thymine dimers and inhibition of DNA synthesis by ultraviolet irradiation of cells.



- Science. 1963;142(3598):1464–66. <https://doi.org/10.1126/science.142.3598.1464>.
35. Shibai A, Takahashi Y, Ishizawa Y, et al. Mutation accumulation under UV radiation in *Escherichia coli*. *Sci Rep*. 2017;7(1):14531. <https://doi.org/10.1038/s41598-017-15008-1>.
  36. Ni P, Huang N, Zhang Z, et al. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics*. 2019;35(22):4586–95. <https://doi.org/10.1093/bioinformatics/btz276>.
  37. Stoiber M, Quick J, Egan R, et al. De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. *Biorxiv*. 2016. <https://doi.org/10.1101/094672>. Accessed 4 August 2025.
  38. Wang X, Ahsan MU, Zhou Y, et al. Transformer-based DNA methylation detection on ionic signals from Oxford Nanopore sequencing data. *Quant Biol*. 2023;11(3):287–96. <https://doi.org/10.15302/J-QB-022-0323>.
  39. An N, Fleming AM, White HS, et al. Nanopore detection of 8-oxoguanine in the human telomere repeat sequence. *ACS Nano*. 2015;9(4):4296–307. <https://doi.org/10.1021/acsnano.5b00722>.
  40. Wiame I, Remy S, Swennen R, et al. Irreversible heat inactivation of DNase I without RNA degradation. *Biotechniques*. 2000;29(2):252–54, 256. <https://doi.org/10.2144/00292bm11>.
  41. Hanaki K, Nakatake H, Yamamoto K, et al. DNase I activity retained after heat inactivation in standard buffer. *Biotechniques*. 2000;29(1):38–40, 42. <https://doi.org/10.2144/00291bm05>.
  42. Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. *arXiv* 2015; <https://doi.org/10.48550/arXiv.1512.04150>.
  43. Ferguson JM, Smith MA. SquiggleKit: a toolkit for manipulating nanopore signal data. *Bioinformatics*. 2019;35(24):5372–73. <https://doi.org/10.1093/bioinformatics/btz586>.
  44. Gamaarachchi H, Samarakoon H, Jenner SP, et al. Fast nanopore sequencing data analysis with SLOW5. *Nat Biotechnol*. 2022;40(7):1026–29. <https://doi.org/10.1038/s41587-021-01147-4>.
  45. Marti H, Borel N, Dean D, et al. Evaluating the antibiotic susceptibility of *Chlamydia*—new approaches for in vitro assays. *Front Microbiol*. 2018;9:1414. <https://doi.org/10.3389/fmicb.2018.01414>.
  46. Hoffmann K, Schott F, Donati M, et al. Prevalence of chlamydial infections in fattening pigs and their influencing factors. *PLoS One*. 2015;10(11):e0143576. <https://doi.org/10.1371/journal.pone.0143576>.
  47. Tan M, Hegemann JH, Sütterlin C, eds. *Chlamydia biology: from genome to disease*. Norwich, UK: Caister Academic Press; 2020. <https://doi.org/10.21775/9781912530281>.
  48. Janssen KJ, Hoebe CJ, Dukers-Muijers NH, et al. Viability-PCR shows that NAAT detects a high proportion of DNA from non-viable *Chlamydia trachomatis*. *PLoS One*. 2016;11(11):e0165920. <https://doi.org/10.1371/journal.pone.0165920>.
  49. Pantchev A, Sting R, Bauerfeind R, et al. Detection of all *Chlamydia* and *Chlamydia* spp. of veterinary interest using species-specific real-time PCR assays. *Comp Immunol Microbiol Infect Dis*. 2010;33(6):473–84. <https://doi.org/10.1016/j.cimid.2009.08.002>.
  50. Rohner L, Marti H, Torgerson P, et al. Prevalence and molecular characterization of *C. pecorum* detected in Swiss fattening pigs. *Vet Microbiol*. 2021;256:109062. <https://doi.org/10.1016/j.vetmic.2021.109062>.
  51. Gutierrez A, Laureti L, Crussard S, et al.  $\beta$ -Lactam antibiotics promote bacterial mutagenesis via an RpoS-mediated reduction in replication fidelity. *Nat Commun*. 2013;4:1610. <https://doi.org/10.1038/ncomms2607>.
  52. McDonald MD, Owusu-Ansah C, Ellenbogen JB, et al. What is microbial dormancy? *Trends Microbiol*. 2024;32(2):142–50. <https://doi.org/10.1016/j.tim.2023.08.006>.
  53. Potgieter M, Bester J, Kell DB, et al. The dormant blood microbiome in chronic, inflammatory diseases. *FEMS Microbiol Rev*. 2015;39(4):567–91. <https://doi.org/10.1093/femsre/fuv013>.
  54. Dorado project. GitHub repository. <https://github.com/nanoporetech/dorado>. Accessed 4 August 2025.
  55. Porechop project. GitHub repository. <https://github.com/rrwick/Porechop>. Accessed 4 August 2025.
  56. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 2019;20(1):257. <https://doi.org/10.1186/s13059-019-1891-0>.
  57. Leonard CA, Schoborg RV, Borel N. Productive and penicillin-stressed *Chlamydia pecorum* infection induces nuclear factor kappa B activation and interleukin-6 secretion in vitro. *Front Cell Infect Microbiol*. 2017;7:180. <https://doi.org/10.3389/fcimb.2017.00180>.
  58. Stary G, Olive A, Radovic-Moreno AF, et al. A mucosal vaccine against *Chlamydia trachomatis* generates two waves of protective memory T cells. *Science*. 2015;348(6241):aaa8205. <https://doi.org/10.1126/science.aaa8205>.
  59. Marti H, Biggel M, Shima K, et al. *Chlamydia suis* displays high transformation capacity with complete cloning vector integration into the chromosomal *rrn-nqrF* plasticity zone. *Microbiol Spectr*. 2023;11(6):e02378–23. <https://doi.org/10.1128/spectrum.02378-23>.
  60. Rayo E, Pesch T, Onorini D, et al. Viability assessment of *Chlamydia trachomatis* in men who have sex with men using molecular and culture methods. *Clin Microbiol Infect*. 2025;31(5):802–7. <https://doi.org/10.1016/j.cmi.2024.12.038>.
  61. Leonard CA, Schoborg RV, Borel N. Damage/danger associated molecular patterns (DAMPs) modulate *Chlamydia pecorum* and *C. trachomatis* serovar E inclusion development in vitro. *PLoS One*. 2015;10(8):e0134943. <https://doi.org/10.1371/journal.pone.0134943>.
  62. Loehrer S, Hagenbuch F, Marti H, et al. Longitudinal study of *Chlamydia pecorum* in a healthy Swiss cattle population. *PLoS One*. 2023;18(12):e0292509. <https://doi.org/10.1371/journal.pone.0292509>.
  63. Squiggle4Viability. Software Heritage Archive. <https://archive.softwareheritage.org/swh:1:snp:6945c472e189ae00bd630a5b15b6314226254954;origin=https://github.com/Genomics4OneHealth/Squiggle4Viability>.
  64. Ürel H, Benassou S, Marti H, et al. Supporting data for “Nanopore- and AI-Empowered Microbial Viability Inference.” *GigaScience Database*. 2025. <https://doi.org/10.5524/102747>.
  65. SquidBase repository for “Nanopore- and AI-empowered microbial viability inference.” 2025. <https://squidbase.org/submission/s/SQB000009>. Accessed 4 August 2025.
  66. Squiggle4Viability project. GitHub repository. <https://github.com/Genomics4OneHealth/Squiggle4Viability>. Accessed 4 August 2025.
  67. DOME-ML annotations for “Nanopore- and AI-empowered microbial viability inference.” 2025. <https://registry.dome-ml.org/rview/66r9zbj2zc>.