Label-free deep learning-based species classification of bacteria imaged by phase-contrast microscopy

Erik Hallström[1], Vinodh Kandavalli[2], Petter Ranefall[1,3], Johan Elf[2], Carolina Wählby[1*]

1 Department of Information Technology, Uppsala University, Uppsala, Sweden

2 Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden

3 Sysmex Astrego AB, Uppsala, Sweden

* carolina.wahlby@it.uu.se

Abstract

Reliable detection and classification of bacteria and other pathogens in the human body, animals, food, and water is crucial for improving and safeguarding public health. For example, the species and its antibiotic susceptibility are essential to treating a bacterial infection. Here we show that phase contrast time-lapse microscopy combined with deep learning is sufficient to discriminate four species of bacteria relevant for human health. The classification is performed on living bacteria and does not require fixation or staining, meaning that the bacterial species can be determined as the bacteria reproduce in a microfluidic device, enabling parallel determination of susceptibility to antibiotics. We evaluate performance of convolutional neural networks and vision transformers, where the best models attain a class-average accuracy of 98%. Our successful proof of principle results suggests that the methods should be challenged with data covering more species and clinically relevant isolates for future clinical use.

Author Summary

Bacterial infections are a leading cause of premature death worldwide, and growing antibiotic resistance is making treatment increasingly challenging. To effectively treat a

patient with a bacterial infection, it is essential to quickly detect and identify the bacterial species and determine its susceptibility to different antibiotics. Prompt and effective treatment is crucial for the patient's survival. A *microfluidic* device functions as a tiny "lab-on-chip" for manipulating and analyzing small amounts of fluids, such as blood or urine samples from patients. Microfluidic chips with chambers and channels have been designed for quickly testing bacterial susceptibility to different antibiotics by analyzing bacterial growth. Identification of bacterial species has previously relied on killing the bacteria and applying species-specific fluorescent probes. We present deep learning models as a fast and cost-effective method for identifying bacteria species directly from phase-contrast microscopy images of living bacteria, at the same time as growth is analyzed. We envision future applications to identify bacterial species in parallel with antibiotics susceptibility testing and thus improve the treatment of bacterial infections.

1 Introduction

This study applies deep-learning methods for species classification of the bacteria *Enterococcus faecalis*, *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa* growing in traps inside a microfluidic chip. Combining deep learning methods for data analysis and microfluidics as a data-generating platform has recently led to significant advances in biotechnology and biomedical research [1]. Moreover, the success of deep learning across various data-driven fields in recent years has led researchers to apply such methods in detecting microbes across various microscopy modalities [2].

The pivotal moment which started the deep learning revolution is commonly accepted to be the development of "AlexNet" by Krizhevsky et al. [3], a deep convolutional neural network (ConvNet) that won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC2012) in 2012 by a large margin [4]. The success of ConvNets is generally attributed to their ability to automatically learn to extract features through sequential processing of the input data [3].

The next few years commenced with several techniques and architecture optimizations that further refined the performance of deep convolutional

networks [5] [6]. The ResNet (Residual Network) [7] architecture and its variants are currently the primary models utilized in the field, enabling the construction of deeper ConvNets with more layers. ResNet was introduced by He et al. [7], winning the ILSVRC competition in 2015.

More recently, Vision Transformers, ViT, a completely novel neural network architecture containing no convolutional filters, has shown to be on par and even surpass convolutional neural networks in various image-processing tasks, including image classification [8]. A transformer is a neural network design initially used in natural language processing for sequence-to-sequence modeling, such as language translation or chat robots, first demonstrated in the "Attention is all you need" paper by Vaswani et al. [9].

Various three-dimensional ResNet variants have been used for video clip classification, fusing temporal and spatial information across frames, yielding promising results [10] [11]. However, more recently, classifiers based on the transformer architecture have been shown to surpass these ConvNet approaches by a large margin [12].

This study uses data from Kandavalli et al. [13], where deep learning methods were applied for segmenting and tracking cells growing in a microfluidic chip imaged by phase contrast microscopy. After each completed time-lapse experiment, Kandavalli et al. applied species-specific Fluorescence In Situ Hybridization (FISH) probes and identified bacterial species using images captured by fluorescence microscopy. This study uses the same type of fluorescence microscopy image data as ground truth. However, we use one fluorescent channel per species and do not apply combinatorial FISH [13].

The overall experimental setup is shown in Fig 1. Fig 1A-B: A mixed species sample is loaded into the microfluidic chip. Fig 1C: A phase-contrast time-lapse is captured, recording the growth and reproduction of bacteria in the traps for one hour with approximately 32 frames. Fig 1C: After fixation and staining, species-specific fluorescent probes attach to each bacteria, and fluorescence microscopy reveals the species. Traps containing only one species are cropped from the phase-contrast timelapse and labeled according to the fluorescent signal. The cropping targets and labels are shown as colored rectangles. Fig 1D: A neural network is trained to classify a single frame or time-lapse of growth in a single trap. Only two out of approximately 200

positions in the microfluidic chip are shown in Fig 1, where half of the positions in an experiment were treated with antibiotics.

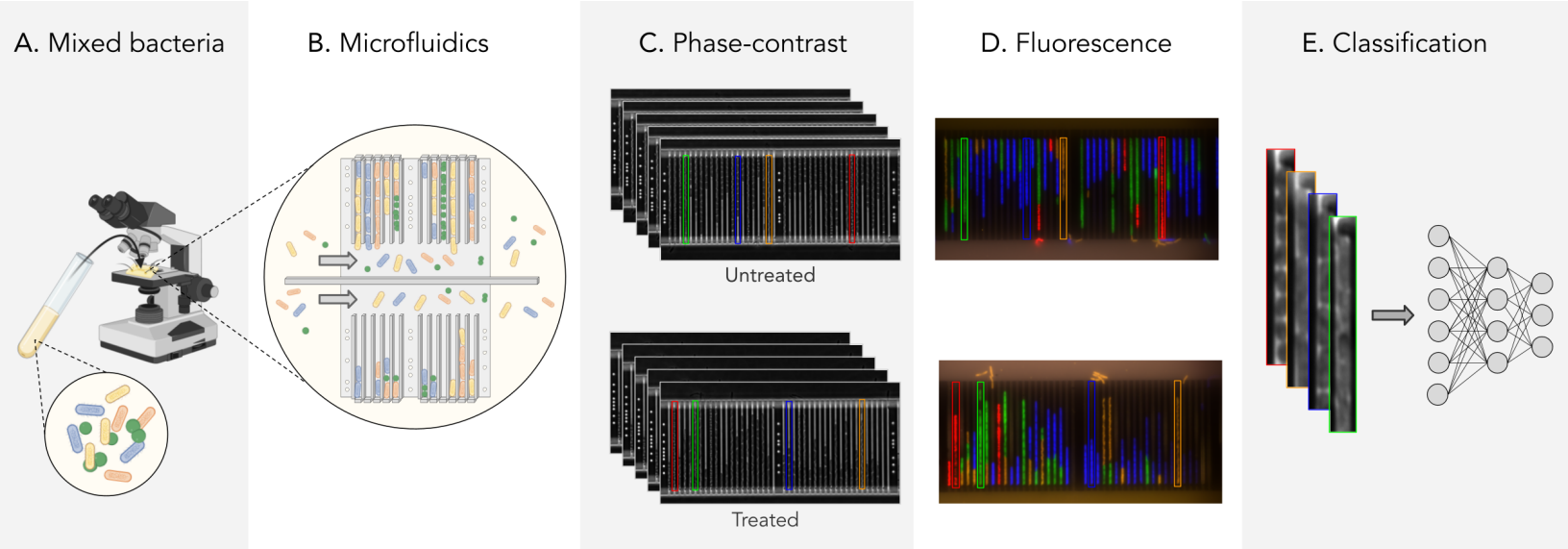


Fig 1. Experiment setup A, B: An illustration of a mixed bacterial sample loaded on a microfluidic chip. C. Time-lapse phase contrast images of bacterial species growing and reproducing in a medium with (bottom) and without (top) antibiotics. D: The correspondence of the fluorescence image to the final phase-contrast image reveals the species, with *Enterococcus faecalis* (red), *Escherichia coli* (green), *Klebsiella pneumoniae* (blue), and *Pseudomonas aeruginosa* (orange). E: Input to the neural network is a single frame or time-lapse video.

To the best of our knowledge, this is the first time ConvNets and Vision Transformers have been used to classify the species of bacteria growing in microfluidic chip traps using only phase-contrast microscopy.

2 Related work

Several methods exist for imaging bacteria, such as bright-field, phase-contrast, fluorescence, and electron microscopy at various magnification levels. In addition to this, various staining, as well as label-free methods, exist. Moreover, spectroscopy and studying images resulting indirectly from scattering patterns are viable methods for bacteria classification. The following is a selection of works in which bacteria classification in micrographs has been performed using convolutional neural networks.

Wang et al. [14] developed a system for detecting and classifying live bacteria growing on agar plates. The device captured image data using coherent microscopy,

scanning the plate every 30 minutes with an image resolution of 4 µm, generating time-lapses. Two deep neural networks were used, one for the early detection of bacterial growth and the second to classify the species of the bacteria using spatiotemporal data. The classification network could classify three species of bacteria with the accuracies 97.2% for *Escherichia coli*, 84.0% for *Klebsiella aerogenes* and 98.5% for *Klebsiella pneumoniae* respectively. A custom convolutional neural network design was used for both tasks, using a Pseudo 3D-network [15] coupled with Dense layers [16].

Hay et al. [17] conducted a study in which fluorescently labeled bacteria inside the larval zebrafish gut were imaged with 3D light sheet fluorescence microscopy. Pixel arrays with suspected bacterial content, possibly containing a bacteria cell, were extracted from this image data and independently labeled by six researchers. The 3D pixel arrays had a size of 28x28x8 pixels, with a resolution of 6.22 px/µm, 6.22 px/µm, and 1.0 px/µm respectively. Finally, a ConvNet, a random forest, and a support vector machine (SVM) were trained to classify the pixel arrays automatically, and the last two methods utilized texture-based feature extraction. The ConvNet reached near-human accuracy and was shown to outperform the other methods with both accuracy and inference speed. Furthermore, transfer learning and augmentation were shown to be highly beneficial for classification accuracy.

Zieliński et al. [18] performed deep learning classification of 33 bacterial species imaged by bright-field microscopy at 100x magnification. The dataset was released as the DIBaS dataset containing 20 images per species. ConvNet backbones from AlexNet, VGG-M, and VGG-VD pre-trained on ImageNet in addition to SIFT descriptors were used for feature extraction. Final species classification was performed using either Support Vector Machines or Random forests. Furthermore, an experiment was conducted where several classifiers were trained by increasing the number of species included, measuring the accuracy as a function of the number of classes admitted. The classifiers based on ConvNet feature extraction were shown to have better accuracy than those using SIFT, and the best models acquired a class-average accuracy at around 96%.

Mai et al. [19] further investigated the DIBaS dataset and developed a more efficient classification ConvNet, better tailored for utilization on resource-limited devices. The network used depth-wise separable convolutions, which consist of a depth-wise convolution with one convolutional filter for each input channel, followed by a

point-wise 1x1 convolution transforming the input to a desired channel depth. The accuracy was measured using 5-fold cross-validation and was shown to be almost on par with Zieliński et al. [18] despite using only 3.2 million parameters, significantly less than the heavy backbones pre-trained on ImageNet. Rotation, shift, shear, scaling, and flip augmentation were shown to be crucial for the performance. Without the augmentations accuracy dropped significantly.

Smith et al. [20] used convolutional neural networks for automated Gram stain classification. Microscopy image data was captured using a 40x dry objective, and the images were then cropped and annotated manually. In total, 100,213 crops were retrieved containing Gram-positive cocci in clusters, Gram-positive cocci in chains/pairs, Gram-negative rods or background. The Inception v3 convolutional neural network pre-trained on ImageNet was applied to the classification task, and the resulting model attained an average classification accuracy of 94.9% on the held-out test crops, consisting of 20% randomly selected samples from the total crops.

Panigrahi et al. [21] used a shape index map calculated from the Hessian of the image as a preprocessing step before feeding the single-channel phase-contrast image data to a U-net for semantic classification of *Myxococcus xanthus* and *E. coli*. Genetically modified bacteria expressing fluorescent markers (GFP and mCherry for *Myxococcus xanthus* and *E. coli* respectively) were used to generate training data. The network was highly accurate despite dense inter-species populations with interactions between the cells. The image data had a resolution of 16.67 px/µm.

The aim of our study presented in this article is to classify bacteria as early as possible, without pre-cultivation or staining. In addition to this, we want to perform species identification in parallel with antibiotics susceptibility testing (AST) on the same microfluidic chip. The ultimate goal is to combine species identification and AST to quickly select suitable patient treatment.

3 Results

The classification accuracy of three types of deep-learning models was evaluated: Vision Transformers (ViT), ResNets, and R(2+1)D "Video ResNets" [10]. The ViT and ResNets were investigated using varying model sizes and sub-sampled input image data.

The Video ResNets were evaluated by feeding the network single or multiple time-lapse frames.

The models were trained and evaluated on 3,396 cropped-out time-lapse videos of single traps with bacteria growing in a microfluidic chip. Each video contained 32 image frames collected at a frame rate of two minutes. The image time-lapse data were collected from several experiments where all four bacterial species were mixed and cultured in the same microfluidic device. In addition to time-lapses of untreated bacteria, the dataset included bacteria treated with different antibiotics (see Materials and Methods). These mixed species datasets were selected to avoid potential classification bias due to chip-to-chip variations. Some traps contained a mix of bacteria from different species, visible from the fluorescence signals captured after the final frame. These traps were excluded from the dataset as the probability of mixed species in the same trap is very low in an actual clinical setting. However, this ground truth selection was occasionally unreliable (see Discussion).

The data were partitioned using random seeds into train/test by trap-basis so that image data from 85% of the traps were used to train the deep learning models, and 15% of the traps were used for testing. All models were trained and tested using the same dataset and identical random seeds.

The accuracy was first calculated separately for each of the four species as the number of correctly classified instances of a particular species, true positives, TP_i over the total number of instances of that species N_i in the dataset, $acc_i = \frac{TP_i}{N_i}$, referred to as class-specific accuracy. The class-specific accuracies were then averaged to obtain the class-average accuracy describing the overall accuracy of a model, $acc_{avg} = \frac{1}{M}\sum_{i=1}^{M} acc_i$, where $M = 4$ is the number of species in the dataset. This is also referred to as the macro-averaged accuracy of a multiclass problem.

3.1 Single-frame classification

We started by evaluating how well our networks could identify bacterial species from the first frame in the time series - this would represent the very first information available in a clinical setting. For this single-frame classification experiment, ResNets and ViTs were trained on all time-frames in the training set but evaluated only on the first frame

of the time-lapse series in the test set. Networks of different sizes were compared, either trained from scratch or pre-trained on ImageNet. The base transformers ViT-B/16 and ViT-B/8 are the standard transformer models from [8] with patch sizes 16 and 8 respectively, the others are downscaled versions of ViT-B/8, with the settings ViT-[d_{embed}]-[h]-[n_{depth}]. As seen in the results in Fig 2 and Fig 3, the pre-trained ViT outperformed the best-achieving ResNet despite the ResNet's inherent inductive bias for image classification. Pre-training was shown to be more critical for ViT than for ResNets, which was expected. Robust gains were made using a deeper network, however, accuracy was saturated when using a ResNet with more than 26 layers. Most classification errors were made on *Pseudomonas aeruginosa*, especially when using smaller models.

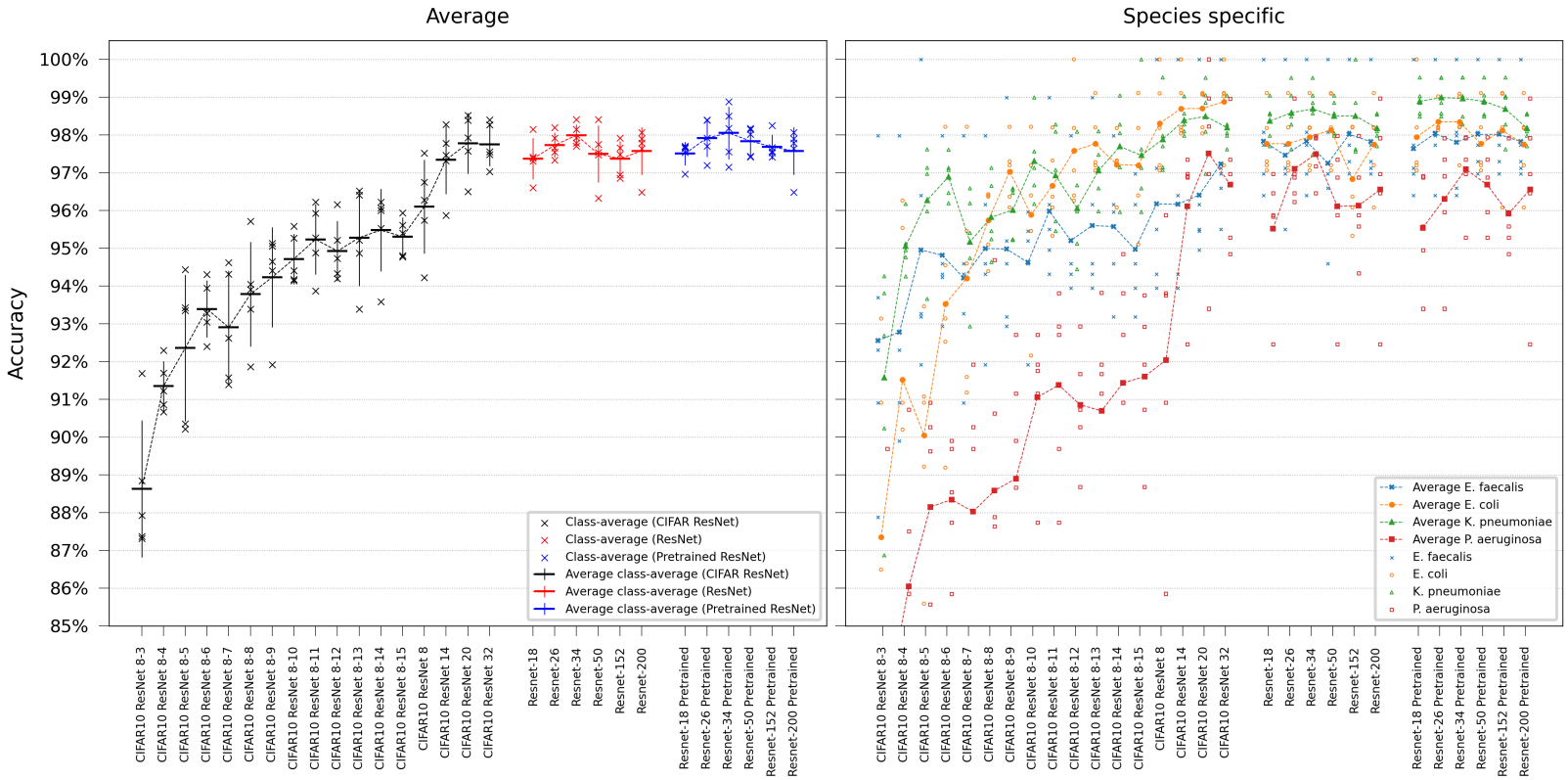


Fig 2. ResNet single-frame classification Model comparison performing single-frame classification of the first frame in the time-lapse using different ResNet variants. The models are grouped after the ResNet family. The error bars show the standard deviation of the class-average accuracy over all five random seeds. The scatter plots show the class-average and class-specific accuracy of all classifiers trained with their respective seeds. A slight jitter was applied along the categorical axis in the species-specific scatter plot to mitigate overplotting. Lines are not for interpolation or statistical inference but are added as a visual guide to track the mean values across the ordinal scale.

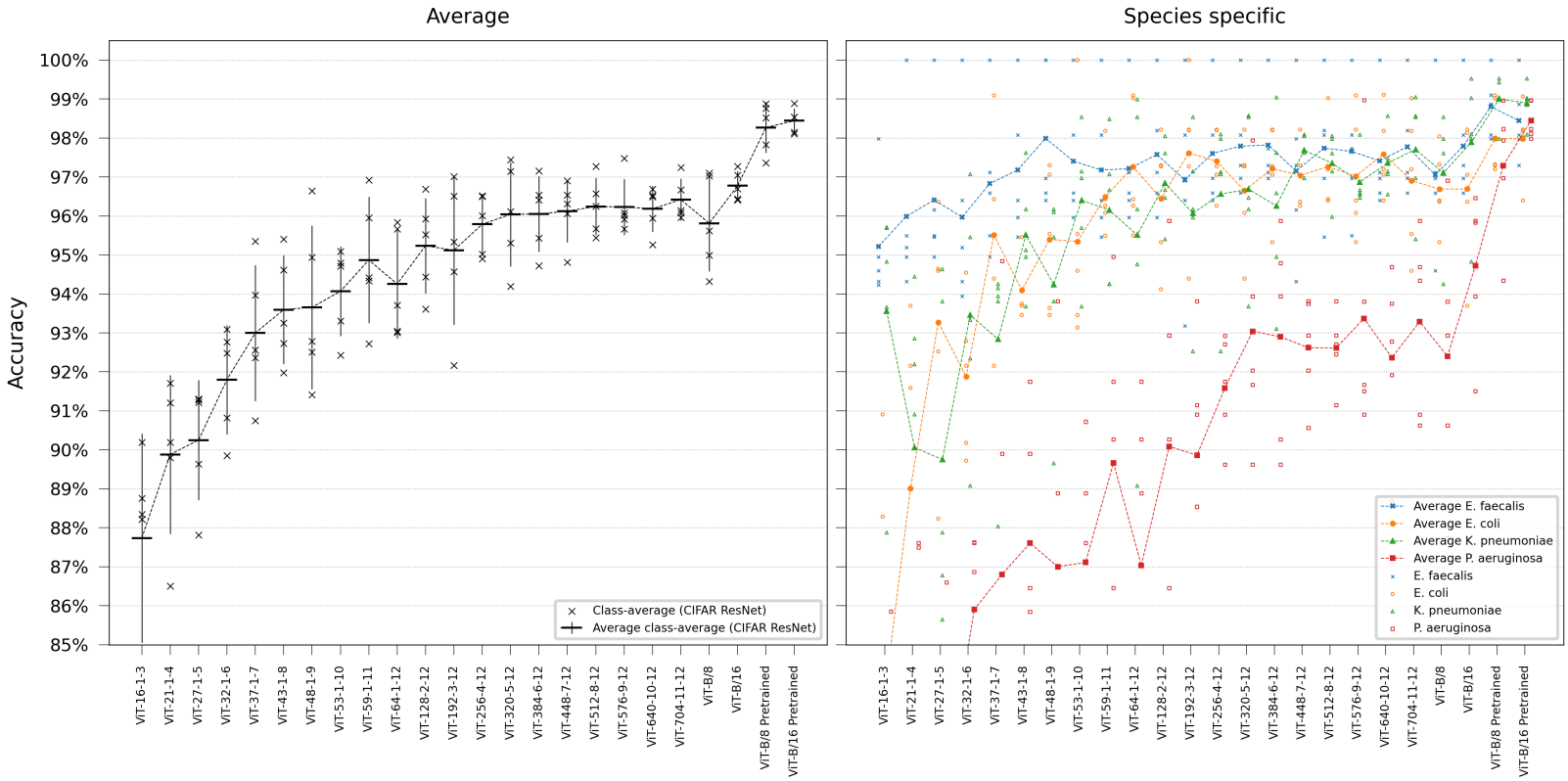


Fig 3. ViT single-frame classification Model comparison performing single-frame classification of the first frame in the time-lapse using different ViT variants. The error bars show the standard deviation of the class-average accuracy over all five random seeds. The scatter plots show the class-average and class-specific accuracy of all classifiers trained with their respective seeds. A slight jitter was applied along the categorical axis in the species-specific scatter plot to mitigate overplotting. Lines are not for interpolation or statistical inference but are added as a visual guide to track the mean values across the ordinal scale.

3.2 Decreasing resolution

Next, experiments were conducted to simulate using lower magnification microscopy, which is essential if the models would be used in a clinical setting, concerning both hardware and computational efficiency. The pre-trained models ResNet-26 and ViT-B/8 were selected for evaluation while training and testing on images with a decreased resolution, sub-sampling images before input from the original image size of 52x382 pixels, step-wise, ending at an image size of 3x22 pixels. The ViT-8 was selected due to being able to handle smaller image sizes than ViT-16, and ResNet-26 was selected since deeper networks did not notably increase accuracy. Since the ViT has a patch size of

8x8 pixels, it was not possible to go to a lower resolution than a width of 8 pixels for the transformer. Therefore, the evaluation was limited so that the smallest images were 8x59 pixels. Sub-sampled images from all image frames of the training dataset were used during training, and testing was performed using a sub-sampled version of the first frame in the hold-out test dataset. As seen in Fig 4 and Fig 5, ViT-B/8 performed better than ResNet-26 when using sub-sampled images.

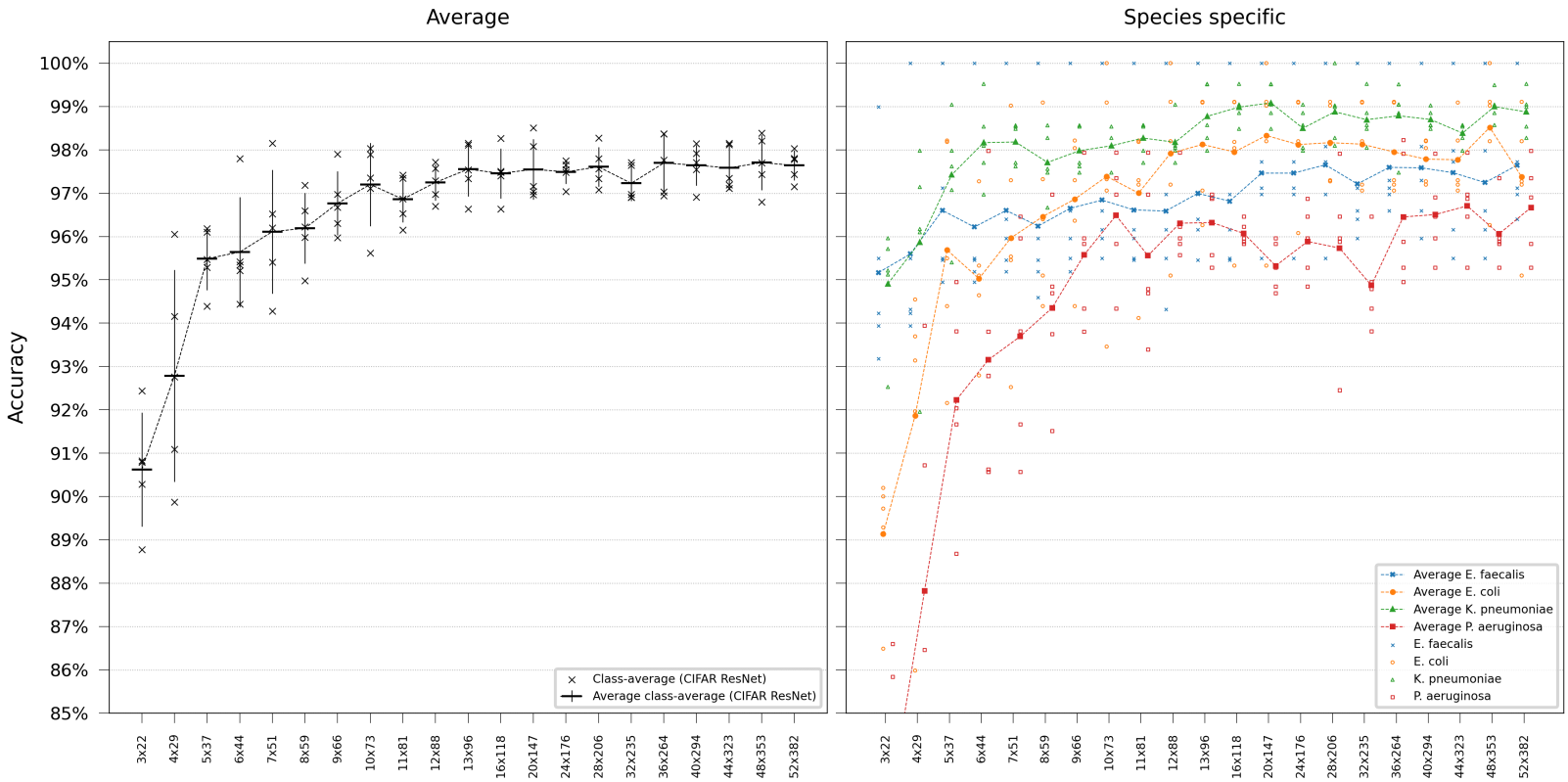


Fig 4. ResNet performance at reduced image resolution. The error bars show the standard deviation of the class-average accuracy over all five random seeds. The scatter plots show the class-average and class-specific accuracy of all classifiers trained with their respective seeds. A slight jitter was applied along the categorical axis in the species-specific scatter plot to mitigate overplotting. Lines are not for interpolation or statistical inference but are added as a visual guide to track the mean values across the ordinal scale.

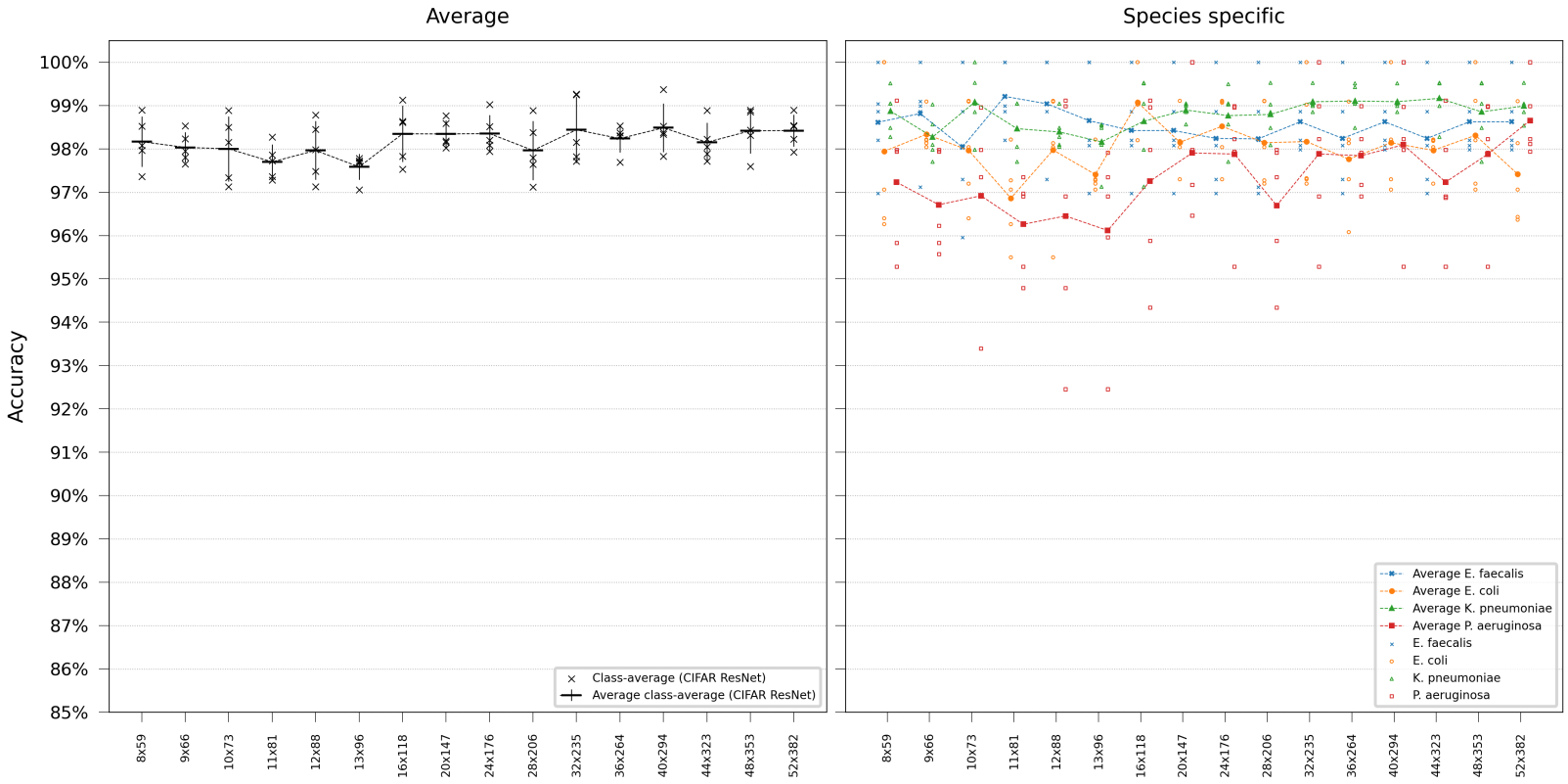


Fig 5. ViT performance at reduced image resolution. The scatter plots show the class-average and class-specific accuracy of all classifiers trained with their respective seeds. A slight jitter was applied along the categorical axis in the species-specific scatter plot to mitigate overplotting. Lines are not for interpolation or statistical inference but are added as a visual guide to track the mean values across the ordinal scale.

3.3 Video classification

After evaluating network performance on single frames, we wanted to explore whether accuracy could be improved using multiple frames and whether the order of the frames was important. The R(2+1)D Video ResNet model was trained using an increasing number of frames as input, with the frames either shuffled or ordered. Testing was performed starting from the first frame in the hold-out test time-lapse dataset. Each model was re-trained five times using a different random seed, thus using different train/test splits. The results in Fig 6 show that accuracy increases with the number of frames used and indicate that maintaining the original order of the image frames leads to a slightly better performance than randomly shuffled frames, especially when using a larger number of frames.

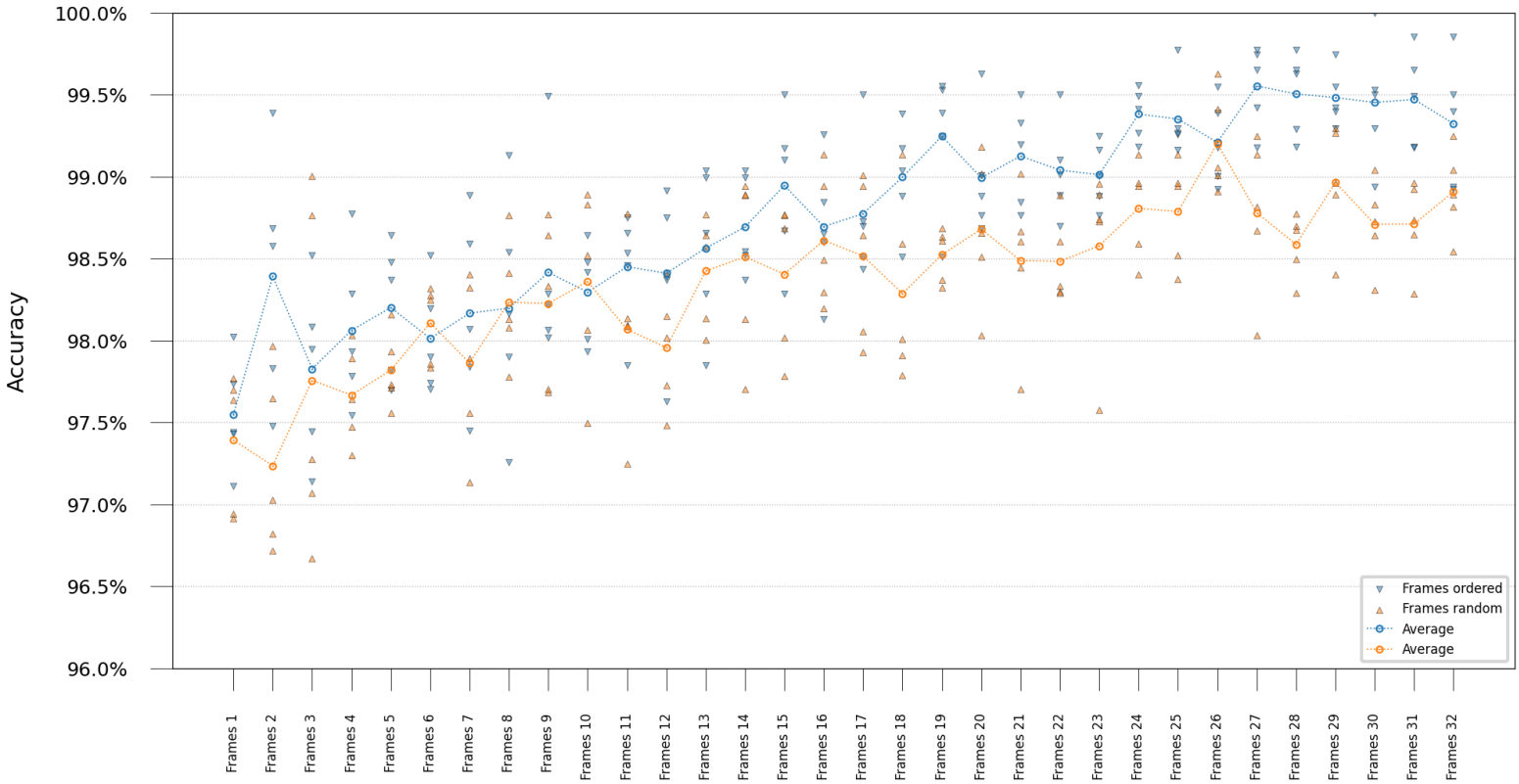


Fig 6. Video ResNet performance at varying number of frames Results from training video classification "R(2+1)D"-networks using various numbers of frames. Ordered and randomly shuffled time-lapses were compared for a fixed number of frames. The scatter plot shows class-average accuracy for each classifier, the line graph shows average class-average accuracy over all classifiers trained with their respective seeds. Lines are not for interpolation or statistical inference but are added as a visual guide to track the mean values across the ordinal scale.

3.4 Model evaluation

A summary of all model results is shown in table 1. The models ResNet-26, ViT-B/8, and 32-frame R(2+1)D Video ResNet from the scaling experiments were selected trained from scratch using an identical random seed. Confusion matrices in Fig 7 show classification errors for these models. In order to evaluate model limitations and analyze image data in instances where the model makes an incorrect assessment, the misclassified test samples from the models in Fig 7 are outlined in S1 Fig - S12 Fig.

ResNet-26 Pretrained

Actual \ Predicted	E. faecalis	E. coli	K. pneumoniae	P. aeruginosa
E. faecalis	99	0	0	0
E. coli	1	100	0	1
K. pneumoniae	0	3	207	0
P. aeruginosa	1	1	1	96

ViT-B/8 Pretrained

Actual \ Predicted	E. faecalis	E. coli	K. pneumoniae	P. aeruginosa
E. faecalis	99	0	0	0
E. coli	0	101	1	0
K. pneumoniae	0	1	209	0
P. aeruginosa	1	0	0	98

R(2+1)D Pretrained

Actual \ Predicted	E. faecalis	E. coli	K. pneumoniae	P. aeruginosa
E. faecalis	99	0	0	0
E. coli	0	102	0	0
K. pneumoniae	0	1	209	0
P. aeruginosa	0	0	0	99

Fig 7. Confusion matrices Confusion matrices of the models trained once from scratch using the same random seed. All models used the full resolution for both training and testing. The Video ResNet "R(2+1)D" used all 32 frames for both training and testing. The single-frame classifiers ResNet-26 and ViT-B/8 trained on all frames in the time-lapse but could only access the first frame during test time.

Table 1. Model results

Video Resnet "R(2+1)D"			
Frames	Parameters	FLOPS	Accuracy
1	31.30 M	15.28 GFlops	$97.55 \pm 0.35\%$
4	31.30 M	32.08 GFlops	$98.06 \pm 0.48\%$
8	31.30 M	60.39 GFlops	$98.20 \pm 0.70\%$
12	31.30 M	92.47 GFlops	$98.41 \pm 0.50\%$
16	31.30 M	120.79 GFlops	$98.70 \pm 0.41\%$
20	31.30 M	152.87 GFlops	$98.99 \pm 0.38\%$
24	31.30 M	181.18 GFlops	$99.38 \pm 0.16\%$
28	31.30 M	213.26 GFlops	$99.51 \pm 0.25\%$
32	31.30 M	241.58 GFlops	$99.32 \pm 0.40\%$

ResNet			
Model	Parameters	FLOPS	Accuracy
CIFAR10 ResNet 8-3	2887	17.32 MFlops	$88.63 \pm 1.82\%$
CIFAR10 ResNet 8-4	5040	29.84 MFlops	$91.35 \pm 0.66\%$
CIFAR10 ResNet 8-5	7789	45.71 MFlops	$92.35 \pm 1.94\%$
CIFAR10 ResNet 8-6	11.13 k	64.96 MFlops	$93.39 \pm 0.75\%$
CIFAR10 ResNet 8-7	15.07 k	87.58 MFlops	$92.90 \pm 1.51\%$
CIFAR10 ResNet 8-8	19.61 k	113.56 MFlops	$93.78 \pm 1.38\%$
CIFAR10 ResNet 8-9	24.75 k	142.91 MFlops	$94.23 \pm 1.33\%$
CIFAR10 ResNet 8-10	30.47 k	175.63 MFlops	$94.72 \pm 0.67\%$
CIFAR10 ResNet 8-11	36.80 k	211.71 MFlops	$95.23 \pm 0.93\%$
CIFAR10 ResNet 8-12	43.72 k	251.17 MFlops	$94.92 \pm 0.80\%$
CIFAR10 ResNet 8-13	51.24 k	293.99 MFlops	$95.28 \pm 1.28\%$
CIFAR10 ResNet 8-14	59.35 k	340.18 MFlops	$95.48 \pm 1.09\%$
CIFAR10 ResNet 8-15	68.06 k	389.74 MFlops	$95.30 \pm 0.51\%$
CIFAR10 ResNet 8	77.36 k	442.66 MFlops	$96.10 \pm 1.24\%$
CIFAR10 ResNet 14	174.58 k	952.83 MFlops	$97.34 \pm 0.90\%$
CIFAR10 ResNet 20	271.80 k	1.46 GFlops	$97.78 \pm 0.81\%$
CIFAR10 ResNet 32	466.23 k	2.48 GFlops	$97.75 \pm 0.57\%$
Resnet-18	11.17 M	1.43 GFlops	$97.37 \pm 0.55\%$
Resnet-26	13.95 M	1.88 GFlops	$97.73 \pm 0.33\%$
Resnet-34	21.28 M	3.00 GFlops	$97.99 \pm 0.28\%$
Resnet-50	23.51 M	3.37 GFlops	$97.50 \pm 0.75\%$
Resnet-152	58.15 M	9.95 GFlops	$97.37 \pm 0.46\%$
Resnet-200	62.63 M	12.71 GFlops	$97.58 \pm 0.63\%$
Resnet-18 pre-trained	11.17 M	1.43 GFlops	$97.50 \pm 0.31\%$
Resnet-26 pre-trained	13.95 M	1.88 GFlops	$97.92 \pm 0.51\%$
Resnet-34 pre-trained	21.28 M	3.00 GFlops	$98.05 \pm 0.70\%$
Resnet-50 pre-trained	23.51 M	3.37 GFlops	$97.84 \pm 0.38\%$
Resnet-152 pre-trained	58.15 M	9.95 GFlops	$97.68 \pm 0.33\%$
Resnet-200 pre-trained	62.63 M	12.71 GFlops	$97.58 \pm 0.63\%$

ViT-B/8 pre-trained - Decrease resolution			
Resolution	Parameters	FLOPS	Accuracy
8x59	85.12 M	1.19 GFlops	$98.16 \pm 0.58\%$
9x66	85.12 M	1.36 GFlops	$98.03 \pm 0.35\%$
10x73	85.12 M	1.53 GFlops	$98.00 \pm 0.75\%$
11x81	85.12 M	1.71 GFlops	$97.70 \pm 0.40\%$
12x88	85.12 M	1.88 GFlops	$97.96 \pm 0.68\%$
13x96	85.12 M	2.05 GFlops	$97.58 \pm 0.30\%$
16x118	85.13 M	4.62 GFlops	$98.35 \pm 0.65\%$
20x147	85.14 M	5.66 GFlops	$98.34 \pm 0.32\%$
24x176	85.16 M	10.52 GFlops	$98.35 \pm 0.43\%$
28x206	85.16 M	12.09 GFlops	$97.96 \pm 0.68\%$
32x235	85.19 M	18.99 GFlops	$98.44 \pm 0.76\%$
36x264	85.20 M	21.13 GFlops	$98.24 \pm 0.32\%$
40x294	85.24 M	29.26 GFlops	$98.49 \pm 0.56\%$
44x323	85.25 M	32.92 GFlops	$98.15 \pm 0.45\%$
48x353	85.29 M	43.15 GFlops	$98.42 \pm 0.52\%$
52x382	85.31 M	47.68 GFlops	$98.42 \pm 0.36\%$

ViT			
Model	Parameters	FLOPS	Accuracy
ViT-16-1-3	15.24 k	19.21 MFlops	$87.73 \pm 2.68\%$
ViT-21-1-4	29.34 k	36.03 MFlops	$89.88 \pm 2.04\%$
ViT-27-1-5	54.60 k	62.80 MFlops	$90.25 \pm 1.54\%$
ViT-32-1-6	87.01 k	95.19 MFlops	$91.80 \pm 1.40\%$
ViT-37-1-7	130.84 k	136.42 MFlops	$93.00 \pm 1.74\%$
ViT-43-1-8	196.47 k	194.10 MFlops	$93.60 \pm 1.40\%$
ViT-48-1-9	270.63 k	257.27 MFlops	$93.66 \pm 2.10\%$
ViT-53-1-10	361.83 k	332.27 MFlops	$94.07 \pm 1.15\%$
ViT-59-1-11	487.82 k	431.43 MFlops	$94.87 \pm 1.62\%$
ViT-64-1-12	621.38 k	534.74 MFlops	$94.25 \pm 1.40\%$
ViT-128-2-12	2.42 M	1.69 GFlops	$95.24 \pm 1.22\%$
ViT-192-3-12	5.40 M	3.48 GFlops	$95.12 \pm 1.91\%$
ViT-256-4-12	9.56 M	5.89 GFlops	$95.79 \pm 0.79\%$
ViT-320-5-12	14.90 M	8.93 GFlops	$96.04 \pm 1.34\%$
ViT-384-6-12	21.42 M	12.59 GFlops	$96.05 \pm 0.97\%$
ViT-448-7-12	29.12 M	16.87 GFlops	$96.12 \pm 0.80\%$
ViT-512-8-12	38.00 M	21.78 GFlops	$96.24 \pm 0.73\%$
ViT-576-9-12	48.06 M	27.32 GFlops	$96.23 \pm 0.72\%$
ViT-640-10-12	59.30 M	33.48 GFlops	$96.18 \pm 0.59\%$
ViT-704-11-12	71.72 M	40.27 GFlops	$96.42 \pm 0.54\%$
ViT-B/8	85.31 M	47.68 GFlops	$95.81 \pm 1.23\%$
ViT-B/16	85.31 M	11.59 GFlops	$96.77 \pm 0.38\%$
ViT-B/8 Pre-trained	85.31 M	47.68 GFlops	$98.27 \pm 0.65\%$
ViT-B/16 Pre-trained	85.31 M	11.59 GFlops	$98.44 \pm 0.32\%$

ResNet-26 Pre-trained - Decrease resolution			
Resolution	Parameters	FLOPS	Accuracy
3x22	13.95 M	38.53 MFlops	$90.62 \pm 1.31\%$
4x29	13.95 M	40.81 MFlops	$92.79 \pm 2.45\%$
5x37	13.95 M	72.41 MFlops	$95.49 \pm 0.72\%$
6x44	13.95 M	75.43 MFlops	$95.64 \pm 1.27\%$
7x51	13.95 M	76.36 MFlops	$96.11 \pm 1.43\%$
8x59	13.95 M	85.43 MFlops	$96.19 \pm 0.82\%$
9x66	13.95 M	106.52 MFlops	$96.77 \pm 0.74\%$
10x73	13.95 M	136.72 MFlops	$97.20 \pm 0.96\%$
11x81	13.95 M	142.23 MFlops	$96.86 \pm 0.54\%$
12x88	13.95 M	153.61 MFlops	$97.25 \pm 0.42\%$
13x96	13.95 M	162.94 MFlops	$97.55 \pm 0.63\%$
16x118	13.95 M	205.98 MFlops	$97.46 \pm 0.58\%$
20x147	13.95 M	351.07 MFlops	$97.55 \pm 0.70\%$
24x176	13.95 M	442.43 MFlops	$97.49 \pm 0.28\%$
28x206	13.95 M	539.01 MFlops	$97.61 \pm 0.46\%$
32x235	13.95 M	647.23 MFlops	$97.23 \pm 0.41\%$
36x264	13.95 M	1.05 GFlops	$97.70 \pm 0.70\%$
40x294	13.95 M	1.19 GFlops	$97.64 \pm 0.47\%$
44x323	13.95 M	1.41 GFlops	$97.58 \pm 0.51\%$
48x353	13.95 M	1.57 GFlops	$97.70 \pm 0.64\%$
52x382	13.95 M	1.88 GFlops	$97.64 \pm 0.35\%$

Test accuracies show average and standard deviation over five runs with different random seeds and, thus, different train/test splits.

4 Discussion

The results suggested that it is possible to train deep learning models to accurately classify the bacteria species *E. faecalis*, *E. coli*, *K. pneumoniae*, and *P. aeruginosa* growing in traps in microfluidic chips based solely on phase-contrast image data. Classification accuracy can be maintained at lower image resolution and, thus, lower computational cost, which enables the models to be implemented in diagnostics tools with limited compute resources. ViTs were shown to outperform ResNets by a small margin, especially at lower resolutions.

All our training and model evaluation relies on defining the true class based on fluorescent staining of the final frame. Although the microfluidic device aims to study the lineage from one mother cell, the traps may occasionally have contained multiple species during growth prior to florescent capture. Some classification errors may therefore be due to artifacts, such as multiple species being present in the trap in the first frame, with one subsequently pushed out. Due to pressure variations, bacteria may also be sucked out of the trap and replaced by another species, or the trap may be empty at the start, with cells being pushed in during the experiment. Furthermore, poor fluorophore absorption by a species may indicate that a trap only contained one species when viewing the fluorescent images, where it, in reality, contained two. Manual inspection of the misclassified traps from the testing samples indicated that the aforementioned had occurred in six out of twelve misclassifications, see S1 Fig - S12 Fig.

For the time-lapse experiments, accuracy increased with the number of frames available to the classifier, and results indicate that the order of the frames may be important for classification accuracy. However, these results are inconclusive since accessing the later frames in the time-lapse ensures higher certainty of the ground truth and is therefore beneficial for classification accuracy; consequently, networks that could access the latter frames may show better results due to this bias.

Our method's inability to handle mixed species in a single trap is a limitation. However, it is expected to be a rare event in a clinical setting.

Due to these previously mentioned issues with the training and test data, it is not conclusive that more accurate results are achieved using multiple frames for these particular species. However, in a future scenario in a clinical setting, accuracy may

increase over time as images are continuously acquired from the traps, providing increased confidence in the species classification. Future development is to include a class when a trap has no bacterial content, is empty, or only contains debris.

Another obvious limitation in this proof of principle study is that we only extend our study to four species. More species and different isolates of the same species encountered in a clinical situation need to be addressed in future work. Further studies should also use pre-trained ViT architectures with lower complexity modified to classify lower dimensions, and explore training and fine-tuning the networks at different resolutions [22].

We envision future applications where diagnostic tools containing microfluidic chips can quickly determine bacterial species and guide efficient treatment.

5 Materials and Methods

5.1 Bacterial strains, sample preparations, and antibiotics

In this study, we used four bacterial strains covering both gram-negative and gram-positive cells namely, *E. coli* K12 MG1655 (DA4201), *K. pneumoniae* (ATCC 13883), *P. aeruginosa* (DA6215) and *E. faecalis* (ATCC 51299). Individual glycerol stocks of four strains were cultured overnight in 5 ml Muller-Hinton (MH) medium at 37°C, then diluted 1:1000 times in a fresh MH medium supplemented with pluronic and allowed to grow for 2 hours at 37°C in a shaking incubator with a speed of 200 rpm. Next, we mix all the strains in an equal concentration and load them in a microfluidic chip. For susceptibility testing, we treat the bacterial strains with different antibiotics such as amoxicillin-clavulanate, ampicillin, ciprofloxacin, doripenem, nitrofurantoin, and gentamicin. The concentrations used here were corresponding to MIC values for the *E.coli* suggested by the European Committee on Antimicrobial Susceptibility Testing (EUCAST).

5.2 Experimental setup and imaging conditions

We used the same experimental setup as in Kandavalli et al. 2022 [23] and Baltekin et al. 2017 [24]. Briefly, after loading the cells in a microfluidic chip cells were exposed to

media with and without antibiotics. For imaging, we used the Nikon Ti2-E inverted microscope equipped with a Plan Apo Lambda 100x oil immersion objective (Nikon). To monitor the growth of the bacteria we captured phase contrast images every two minutes for an hour. After time-lapse imaging, the bacteria species in each trap in the microfluidic chip was identified as described in the genotyping protocol in Kandavalli et al 2022. Next, we capture fluorescence images for each probe in four different fluorescence channels (Alexa Fluor 488, Cy3, Cy5, and Texas Red) corresponding to *E. faecalis*, *E. coli*, *K. pneumonia*, and *P. aeruginosa*, respectively.

5.3 Data selection

The microfluidic chips used in the experiments contained 34-44 traps at each capture position and were imaged at 162-180 capture positions depending on the experiment. Binary barcode labels were evenly laid out so that a unique identifier address could be assigned to each trap, as seen in Fig 1. Each phase-contrast and fluorescence microscopy frame was imaged with an image size of 1824 x 3888 pixels at 33 pixels/µm resolution.

In order to obtain the data for this study, a custom image processing software was developed to find and crop each growth trap and barcode label from the microscopy output. The phase-contrast stack was stabilized using rigid body transformations [25] using only the barcode labels as input. Furthermore, this image processing pipeline contained a procedure to shift the position of the fluorescence images to align them with the stabilized phase-contrast time-lapse stack. This procedure used a vertical projection of the pixel intensities of the different images into a one-dimensional vector, generating peaks at the horizontal location of the traps. The peaks from the fluorescence images and the phase-contrast projections were then shifted until alignment, and this shift was then repeated in the original images. The images of the cropped-out traps in the microfulidic chip had a width of 50-54 pixels and a height of 1400-1500 pixels. The cropping software and code for reproducing the deep-learning experiments are freely released along with the dataset to facilitate further research [26].

Traps containing only one species in the mixed species experiments were manually selected by inspecting the final fluorescence image. This task was performed in a ”semi-automated” way by sorting the traps by aggregated statistics of fluorescence pixel

intensity values and visualizing the fluorescence channels side by side, noting the traps with only one bacteria species. A total of 3396 traps with corresponding time-lapses imaging single-species bacterial growth were extracted from seven different experiments divided into 684 *E. faecalis*, 770 *E. coli*, 1280 *K. pneumoniae*, and 662 *P. aeruginosa*. A number of train/test splits were created using predetermined random seeds, withholding 15% of the samples for testing. The dataset contained traps with and without antibiotic treatment during growth, as seen in Table 2.

Table 2. Data selection

	Enterococcus faecalis		*Escherichia coli*		*Klebsiella pneumoniae*		*Pseudomonas aeruginosa*		All species
	Treated	Untreated	Treated	Untreated	Treated	Untreated	Treated	Untreated	Total
Experiment 1	11	12	51	41	3	261	132	90	601
Experiment 2	38	60	0	1	6	183	90	60	438
Experiment 3	51	91	20	116	0	191	2	12	483
Experiment 4	108	66	67	105	0	201	4	4	555
Experiment 5	15	15	35	72	1	142	74	49	403
Experiment 6	6	78	7	41	0	1	1	5	139
Experiment 7	83	50	8	206	0	291	109	30	777
All experiments	312	372	188	582	10	1270	412	250	3396

Data selection across all seven experiments. Each entry corresponds to a microfluidic trap containing only one species of bacteria.

The reason for using mixed species across several experiments was to mitigate the possibility of overfitting to experimental settings, making the model base the decision on spurious correlation in the data, the so-called "Clever Hans"-prediction [27], perhaps based on microscopy settings, background features or overinterpretation and adapting to non-salient features [28].

5.4 Single frame classification

When evaluating network performance on single frames, image classifiers were trained using only a single frame as input. At test-time, the networks could only access the first frame in the time-lapse since it is clinically more relevant to obtain the correct species of the bacteria as early as possible. During training, all frames were utilized. We compared the results of multiple ResNet architectures with the novel ViTs, both with and without transfer learning. Although various modifications and improvements to the ViT and ResNet architecture exist, we tested the vanilla versions to investigate each architecture's inherent potential. Recent findings suggest that training regimes, augmentation, and hyper-parameters can impact performance more than model

architecture [29].

5.4.1 ResNets - Convolutional Neural Networks

In this paper, 21 different ResNet [7] architectures were evaluated; the regular ImageNet ResNet with 18, 26, 34, 162, and 200 layers, as well as the smaller CIFAR-10 ResNet with 8, 14, 20, and 32 layers. In addition to this, ablation experiments with the CIFAR ResNet-8 were conducted, reducing the base feature channel depth in the first layer to r, $r = \{3, ..., 16\}$, named ResNet-8-r. This scaling strategy implied that we used feature channel depths r, $2r$, and $4r$ in the network respectively. ResNet-8-16 is equivalent to CIFAR ResNet-8 since 16 is the default channel depth after the first layer.

The reason for using the smaller CIFAR ResNet variant was twofold; firstly, to determine whether this is a trivial problem or if robust gains can be obtained by using a deeper model with more parameters, secondly to assess the limitations if the models were to be employed in a diagnostics tool with limited compute resources for use in hospitals.

The ResNet ImageNet variant used in our experiments was aimed at ImageNet-classification [4] in the original paper by He et al. [7]. It employed an initial set of 64 7x7 convolutional filters, and a max-pooling downsampling followed by a series of layers using 3x3 filters, grouped in pairs, referred to as "blocks". Convolutions with a stride of 2 in the initial layer of an evenly spaced subset of these blocks were used, doubling feature-channel depth and halving spatial dimensions. Additive skip connections were introduced around each block. In instances where there was a discrepancy in channel depth, it was corrected by using 1x1 convolutions, transforming the feature maps to the same depth. Bottleneck blocks were used in place of the standard two-layer blocks for the deeper networks, allowing for less computationally expensive processing of the feature maps. The bottleneck blocks consisted of three layers; 1x1 filters reducing the channel depth, 3x3 filters processing at the lower channel depth, and finally, 1x1 filters transforming the feature map to the original dimension before the depth reduction.

The CIFAR10 ResNet variant used in our experiments was presented in the original paper by He et al. [7] and aimed to classify the smaller CIFAR10 dataset, consisting of 32x32 color images. It utilized a single 3x3 convolutional layer followed by the standard

blocks but increased feature map depths to 16, 32, and 64, respectively, in the first layer at evenly spaced blocks. All ResNet variants included a global spatial average pooling layer, a fully connected layer, and a softmax activation function as a final layer. Batch normalization [6] was used before each activation.

5.4.2 ViT - Vision Transformers

We evaluated 25 variants of the ViT architectures, the regular ViT-B with pre-trained and randomly initialized weights using patch sizes 8 and 16, and downscaled versions of ViT-B/8. The scaling was performed as follows; the number of heads and the embedding dimension was linearly scaled down, keeping the number of layers constant at 12, analogously to the architectures presented in DeiT [30], ending up using single-head attention and an embedding dimension of 64. This scaling strategy was followed by proportionally scaling down the number of layers and embedding dimension, ending up in an embedding dimension of 16 and a depth of 3 layers, reducing the model complexity and observing the average accuracy. We name these architectures ViT-[d_{embed}]-[h]-[n_{depth}], where ViT-B/8 being equivalent with ViT-768-12-12.

In the ViT architecture [8], the input image was first split into N patches with a predefined size, and the patches were then flattened and linearly embedded into tokens with dimensionality d_{embed}. A learned positional encoding was added to each token based on its position in the sequence. Each of the tokens was then projected into queries, keys, and values with dimensionality d_q, d_k, and d_v using a set of projection matrices with learnable weights $W_i^Q \in \mathbb{R}^{d_{embed} \times d_q}$, $W_i^K \in \mathbb{R}^{d_{embed} \times d_k}$, and $W_i^V \in \mathbb{R}^{d_{embed} \times d_v}$. The projection was executed several times in parallel, called "heads" h, $\forall i \in \{1, ..., h\}$. The queries, keys, and values were then packed as rows into matrices Q_i, K_i, and V_i, followed by an operation referred to as the self-attention mechanism.

$$Attention_i(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i$$

Where $Attention_i(Q_i, K_i, V_i) \in \mathbb{R}^{N \times d_v}$. The outputs from all of the h heads were then concatenated column-wise into a matrix $O \in \mathbb{R}^{N \times h d_v}$ and then transformed back to the original dimensionality by a learnable matrix $W^O \in \mathbb{R}^{h d_v \times d_{embed}}$. Dimensionalities $d_q = d_k = d_v = d_{embed}/h$ were chosen for the projection matrices to

make the computational cost of the multi-head attention similar to single-head attention with full dimensionality. Additive skip-connections were added to this multi-head attention computation's output, as well as a layer-wise normalization across all embeddings. A fully connected 2-layer feed-forward neural network then processed each embedding, with $4d_{embed}$ hidden units, containing a skip connection and a layer-wise normalization. These computations were grouped into a block that was repeated n_{depth} times. A learnable "classification token" was prepended to the patch embeddings and fed simultaneously through the network. The final output was based on this token being fed into a classification head consisting of a multilayer perceptron using a single layer at fine-tuning time.

5.4.3 Decreasing resolution

Finally, the performance of ConvNets and ViTs were compared when operating under decreased resolution. In order to approximately simulate using lower magnification microscopy, the Lanczos [31] interpolation was used for downsampling, which approximates the Sinc function. A convolution by a Sinc is equivalent to a low-pass filter in the frequency domain, removing high-frequency parts, thus effectively reducing the resolution. The image data was resized to a width of w, $w \in \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52\}$. A total of 21 pre-trained ResNet-26 networks and 16 ViT-B/8 networks were trained using the subsampled image data at both train and test time.

5.5 Time-lapse classification and temporal features

To investigate whether temporal dynamics were essential for classification accuracy, several "R(2+1)D"- Video ResNet classification networks introduced by Tran et al. [10] were trained on the time-lapse data. A total of 32 frames were obtained from each trap, duplicating the last frame if the number of frames was lower than 32 in the time-lapse and cropping the time-lapse at 32 frames if it contained more than 32 frames. In total, 32 networks were trained using different numbers of frames as input, $n_{frames} \in \{1, 2, ..., 32\}$. Afterwards, identical video classification networks were trained but randomly shuffling the frames before input. The networks trained on time-lapses

with fewer than 32 frames utilized temporal jittering, [10] randomly sampling n_{frames} consecutive frames from the time-lapse during training.

Testing was performed starting from the first frame in the time-lapse. We hypothesize that if temporal information mattered, the networks processing shuffled frames would consistently yield a lower accuracy compared to the networks trained on frames in the correct order.

The R(2+1)D Video ResNet network used repeated blocks with 2D spatial convolutions, processing each frame separately, followed by 1D temporal convolutions, fusing temporal information across frames [10]. The networks were pre-trained on the Kinetics-400 dataset [32] and fine-tuned on our bacterial dataset.

5.6 Training regimes

The deep learning models were all trained with stochastic gradient descent using the AdamW [5] method, applying standard augmentations such as random crops, horizontal flips, randomly shifting, scaling, and rotating, as well as randomly chaining the brightness and contrast of the image. A cosine learning rate scheduler with warm restarts [33] was used, using five warm-up epochs with a decay rate of 0.5. The cycles had 100 epochs each for the single-frame classification networks and 75 for the video classification networks.

For the video classification, R(2+1)D Video ResNet networks with a batch size of 8 and a learning rate of 0.0002 was used. The single frame classification used a batch size of 32. ResNets had a learning rate of 0.001, and the ViTs required a lower learning rate of 0.0001 to converge. The learning rate of the 34-layer ResNet was reduced to 0.0001 due to problems of converging using the standard learning rate. The Video ResNets were trained for 150 epochs, single-frame ResNets for 300 epochs, and the ViT for 500 epochs. The ViT architecture lacks the inherent inductive bias of ResNet and thus required more epochs to converge. A random weighted sampler was used to equally sample the data and mitigate the small class imbalance. Each cropped trap had a width of 52 pixels, after the random crop it had a size of 52x382 pixels.

All networks were retrained five consecutive times with different predetermined random seeds, resulting in different augmentation sequences and train/test splits. The

mean and standard deviation of the average accuracies of the classifications were calculated and presented in the results section.

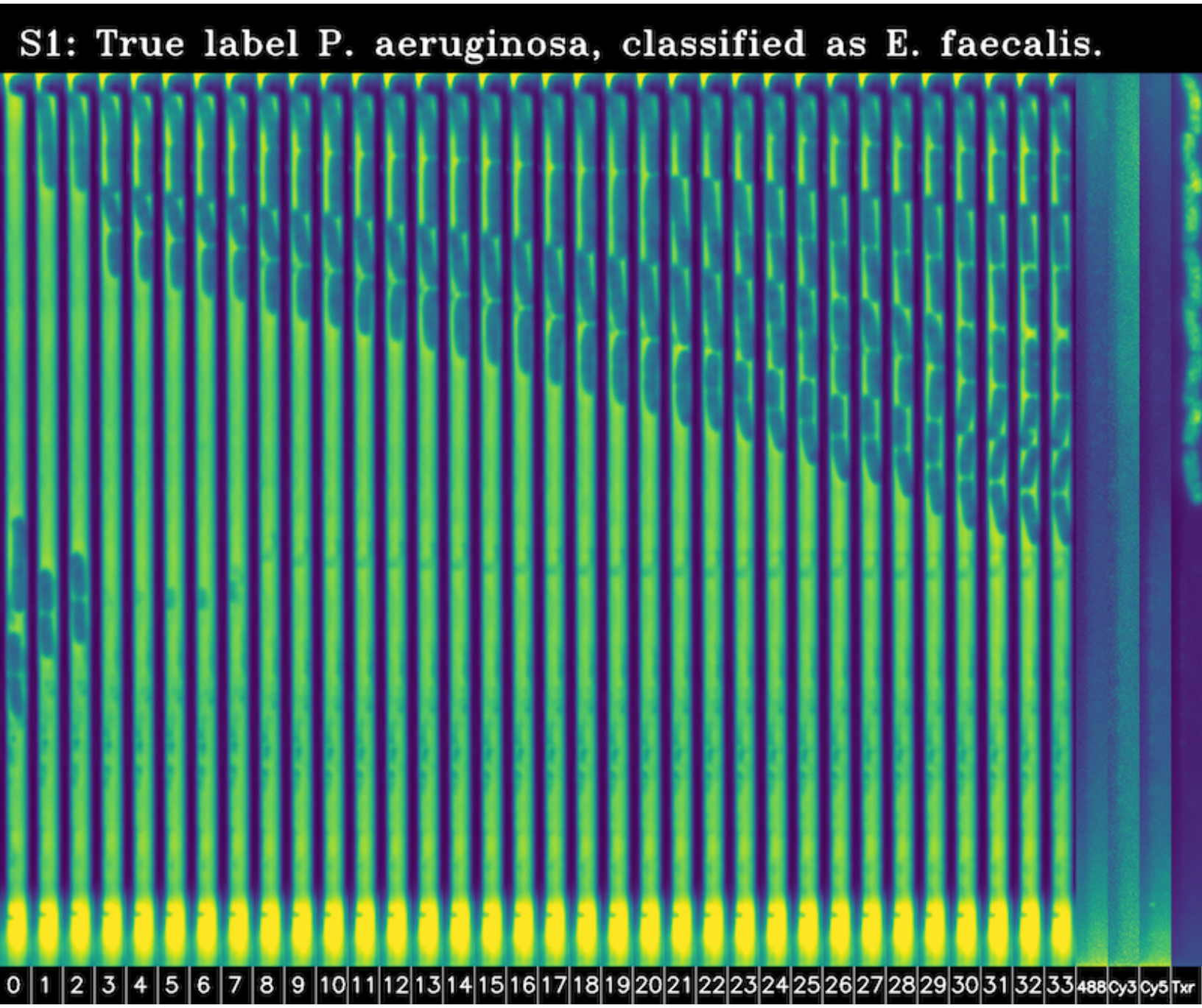
The pre-trained networks were pre-trained on 3-channel natural images, on the contrary, the phase-contrast image data had only a single channel as input. The pre-trained filters in the first layer with filter size n x n x 3 were reshaped to n x n x 1, computing the average across the last axis to accommodate for this. The open-source libraries PyTorch [34] and PyTorch Image Models [35] were used, for full details refer to the released software and dataset [26].

6 Acknowledgments

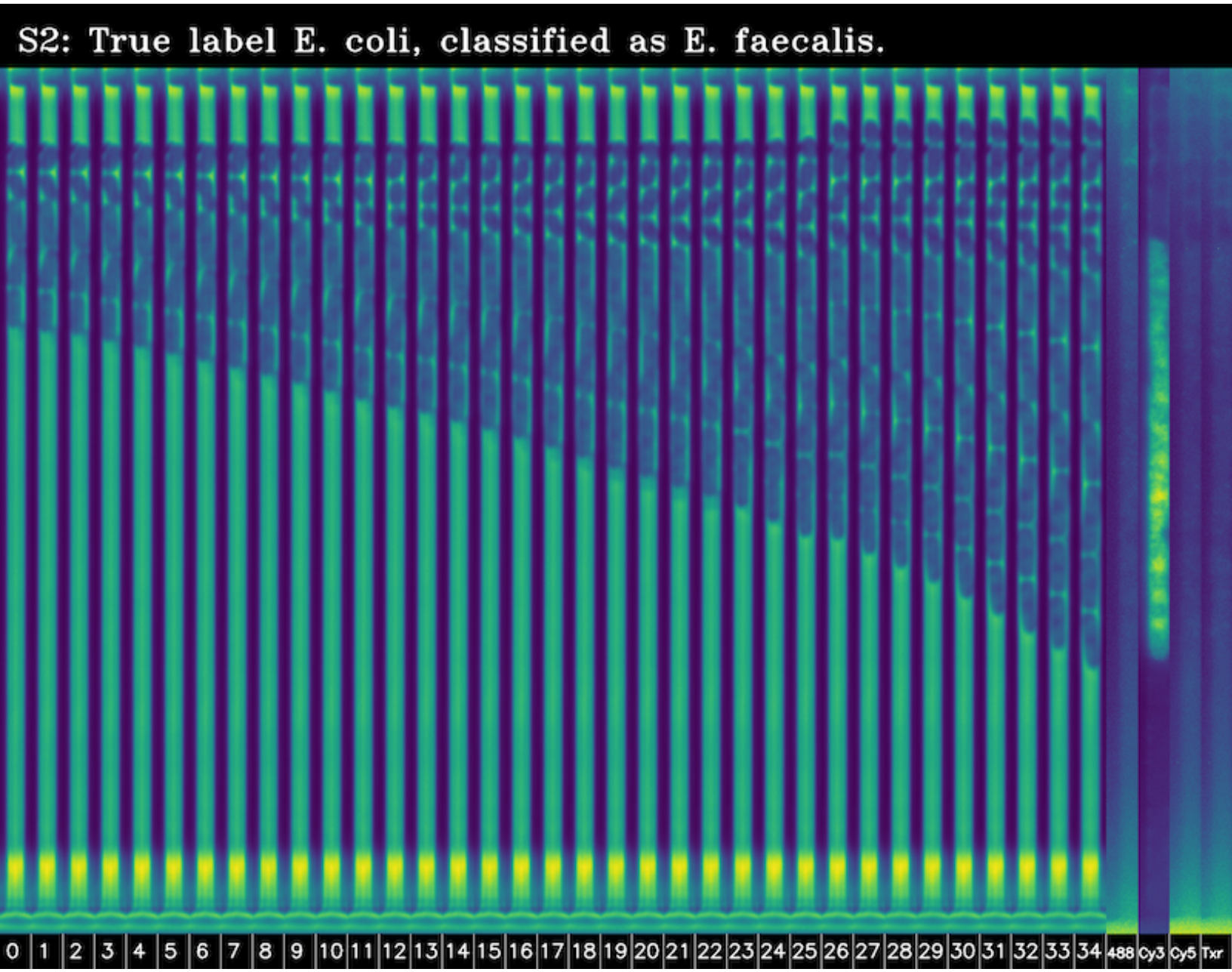
We acknowledge funding from the Swedish Foundation for Strategic Research, SSF ARC19-0016. The computations were enabled by the supercomputing resource Berzelius provided by the National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg foundation. We would also like to thank Anders Hast for valuable input on the manuscript. Johan Elf has founded and is partly engaged in the AMR diagnostics company Sysmex Astrego AB.

7 Supporting information

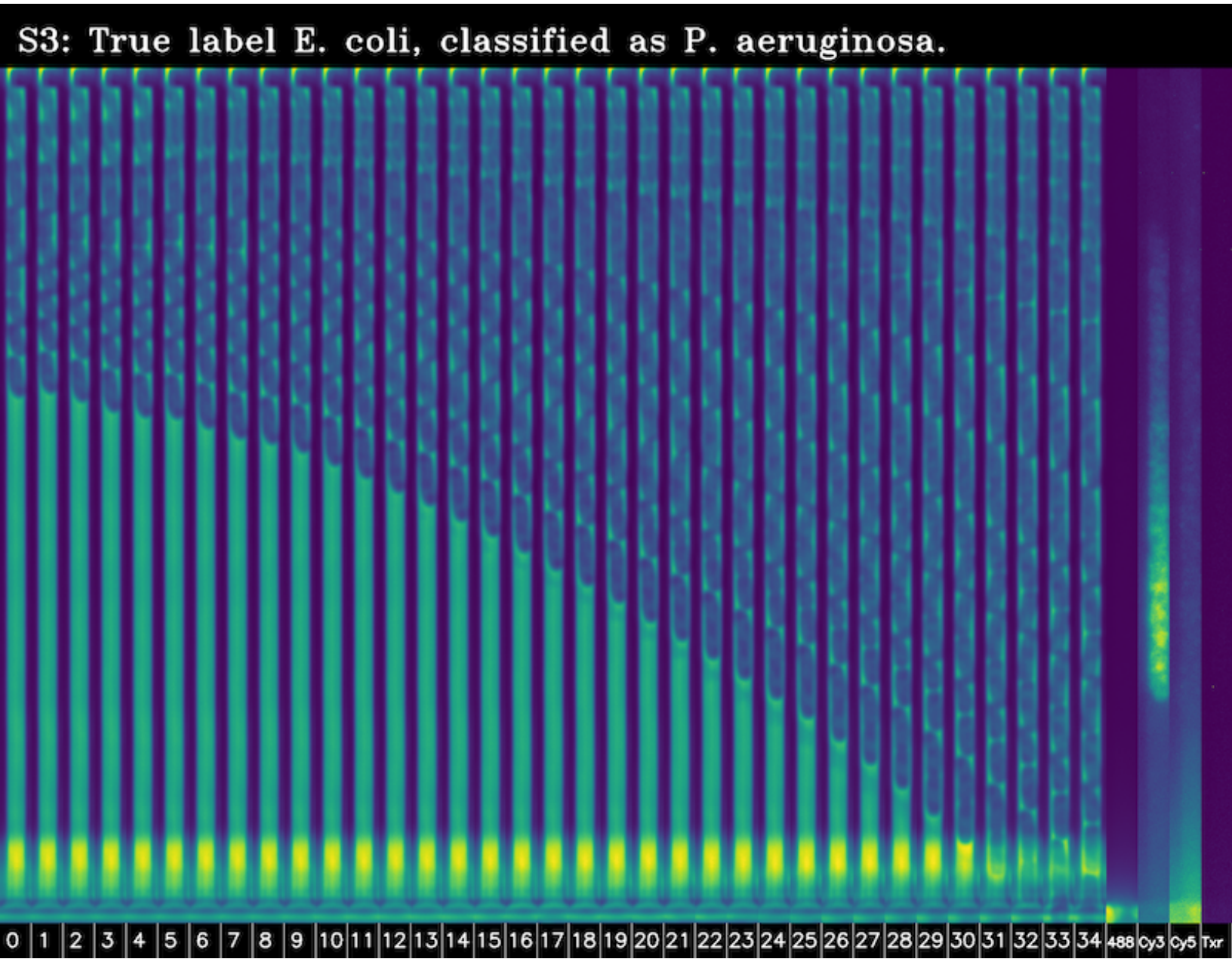
The following are misclassifications made by the models in 7, fine-tuned on the bacteria dataset and tested from the first frame. Labels N, $N \in \{0, ..., 35\}$ show ordered phase-contrast image data for a single trap in the time-lapse, labels L, $L \in \{A, B, C, D\}$ show the final fluorescent imaging after the last frame showing probes Alexa488, Cy3, Cy5, and Txr (Texas Red) respectively, corresponding to species *E. faecalis*, *E. coli*, *K. pneumoniae*, and *P. aeruginosa*.



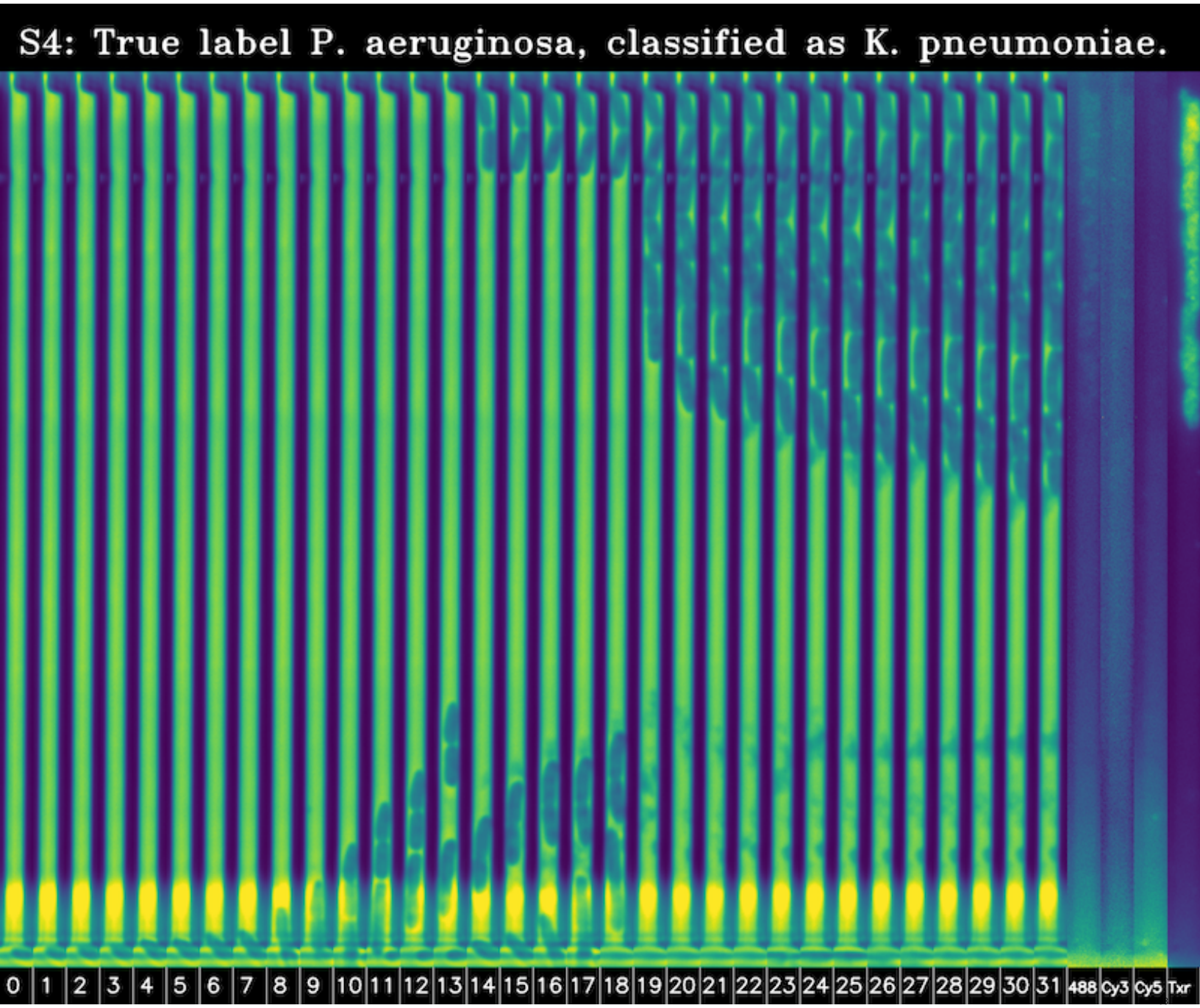
S1 Fig. Misclassified by ResNet-26 True label *P. aeruginosa*, classified as *E. faecalis*. The ResNet may have confused the stop at the top of the trap as a coccus. The upper half of the trap was empty in the first frame.



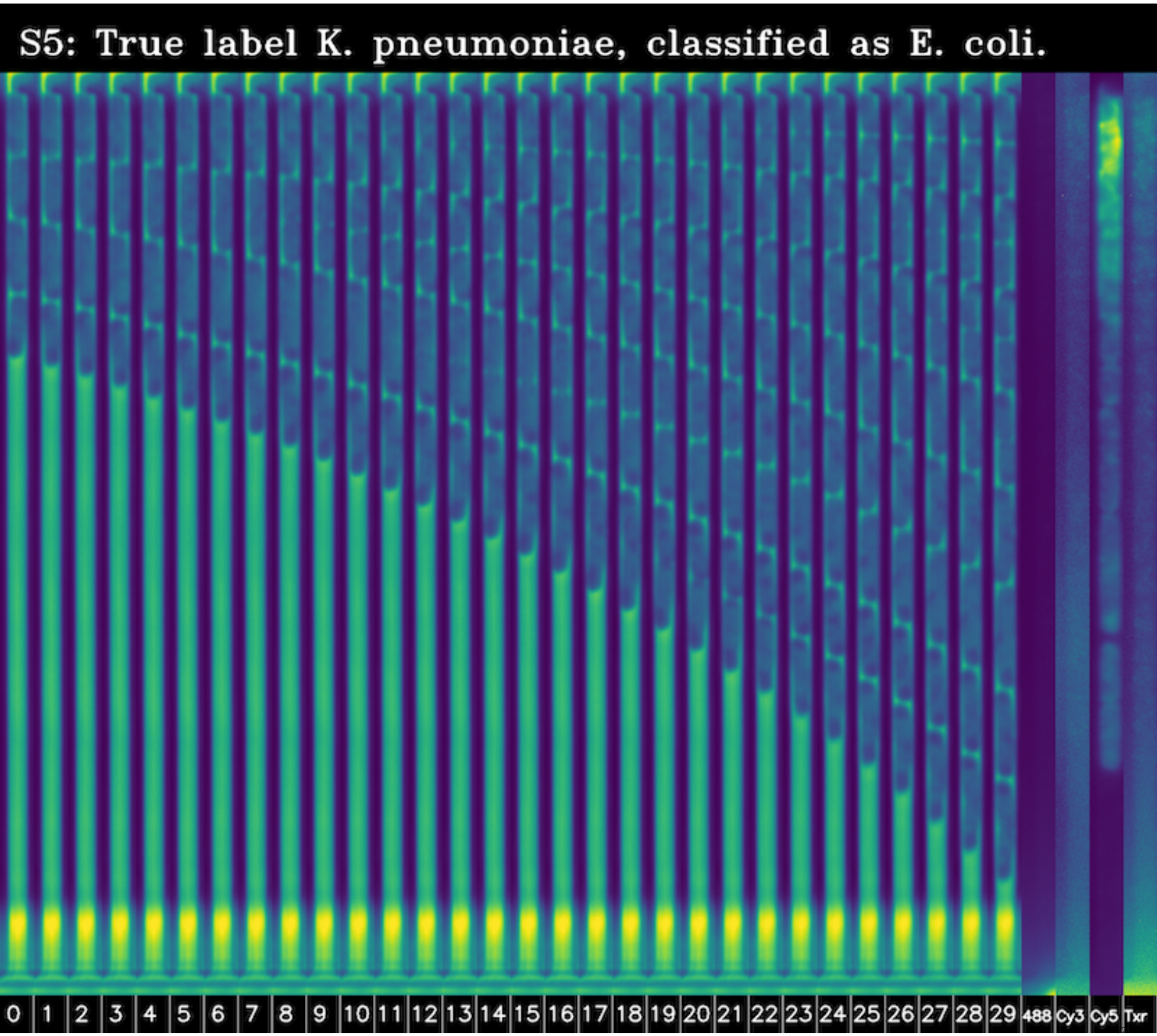
S2 Fig. Misclassified by ResNet-26 True label *E. coli*, classified as *E. faecalis*. It appears to have been cocci in the trap that avoided staining, and the trap clearly did not only contain a single bacterial species.



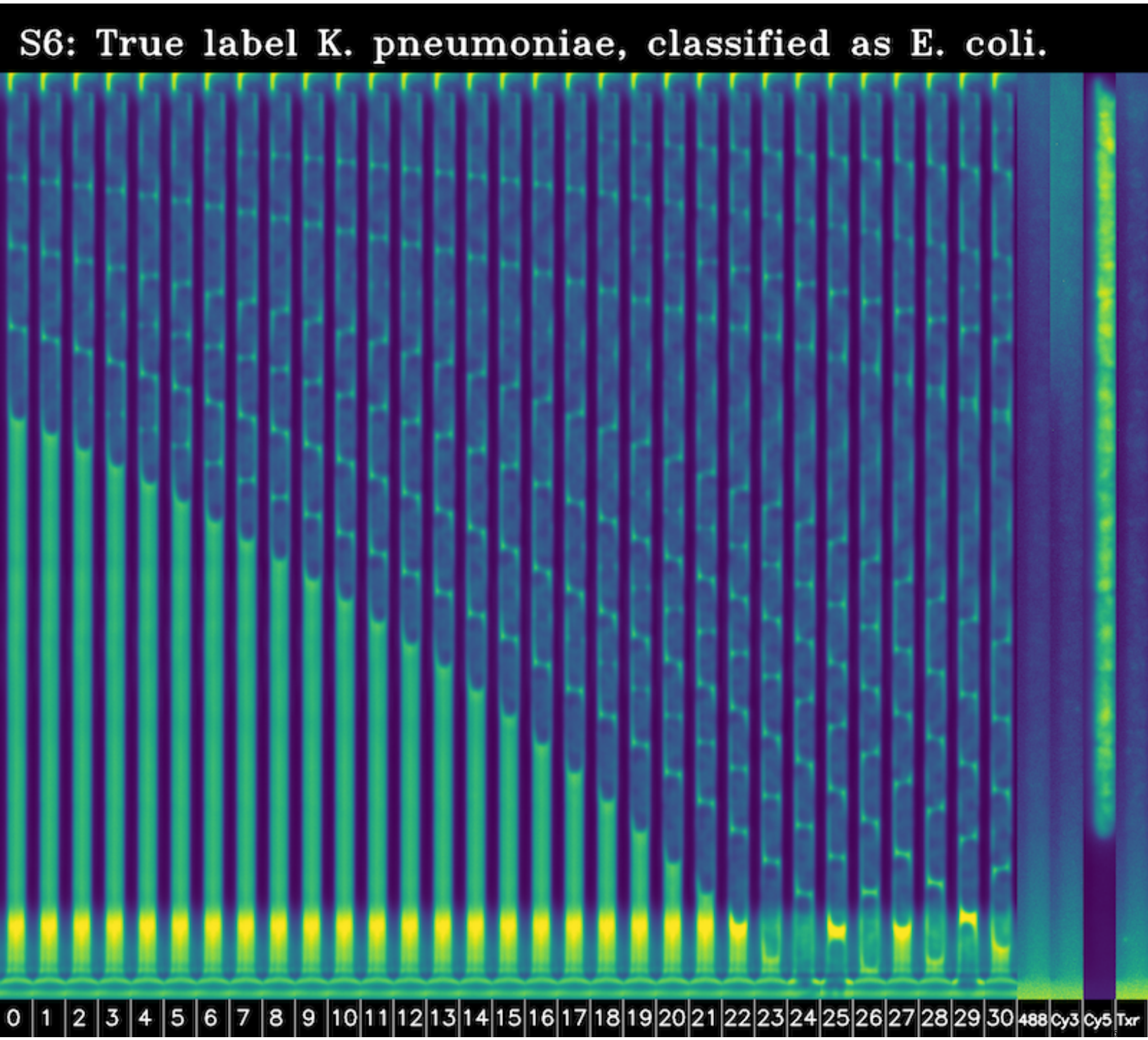
S3 Fig. Misclassified by ResNet-26 *E. coli*, classified as P. aeruginosa. Both species are rods with similar shapes and are easily confused, and the fluorescent staining indicates that there may be two bacterial species in the trap, where one has avoided staining.



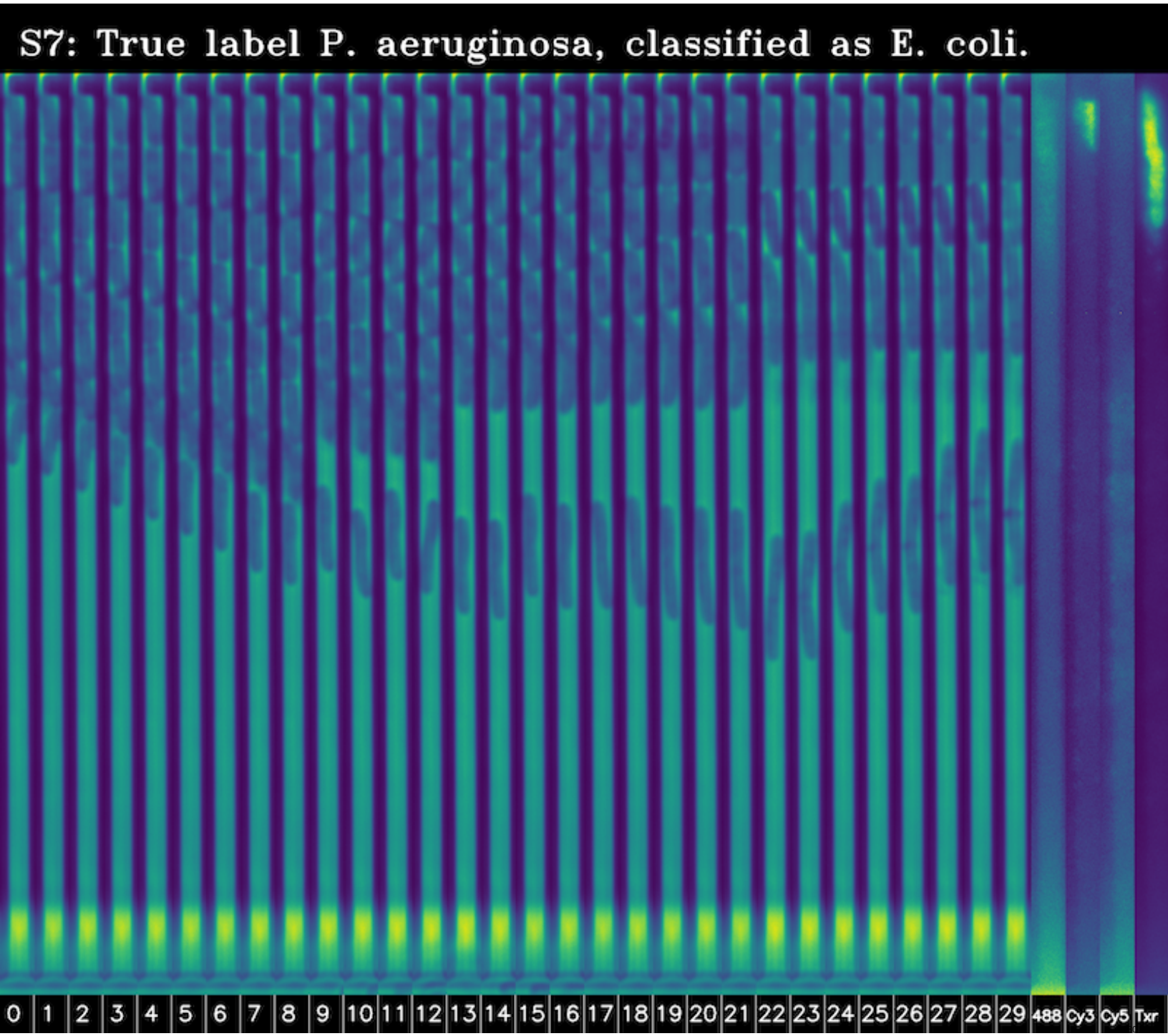
S4 Fig. Misclassified by ResNet-26 True label *P. aeruginosa*, classified as *K. pneumoniae*. The trap was empty in the first frames.



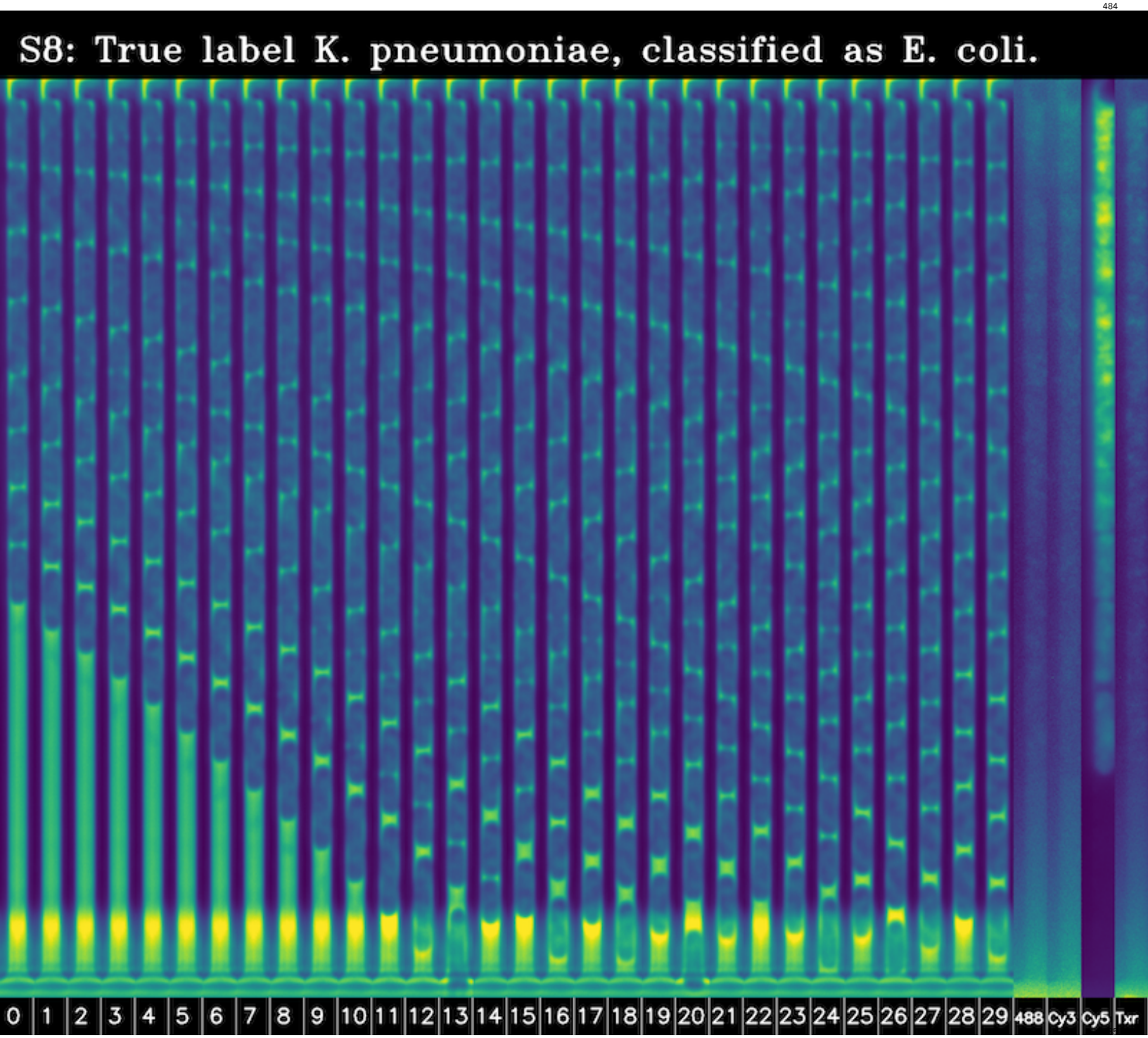
S5 Fig. Misclassified by ResNet-26 True label *K. pneumoniae*, classified as *E. coli*. Both species are rods with similar shapes and are easily confused.



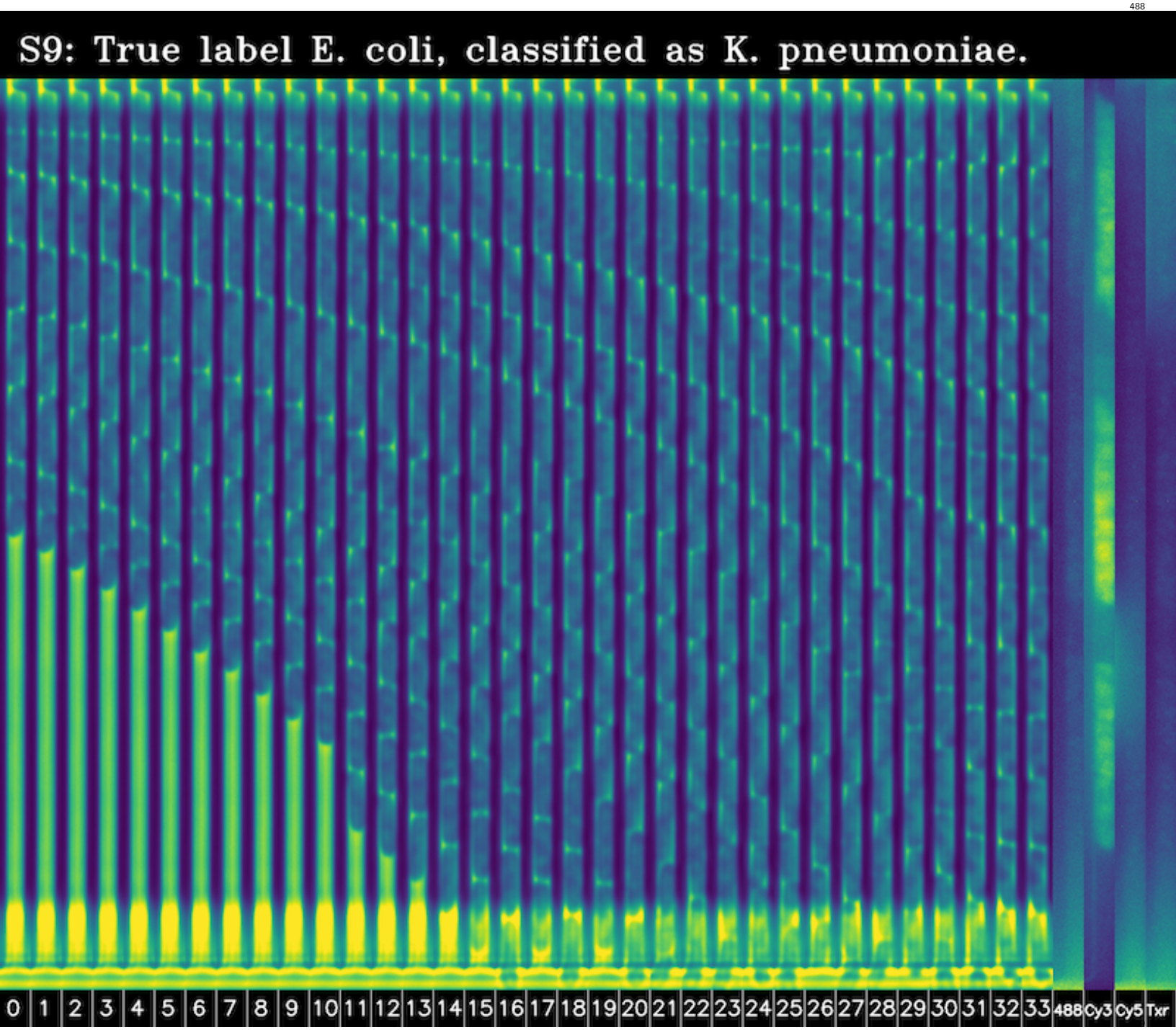
S6 Fig. Misclassified by ResNet-26 True label *K. pneumoniae*, classified as *E. coli*. Both species are rods with similar shapes and are easily confused.



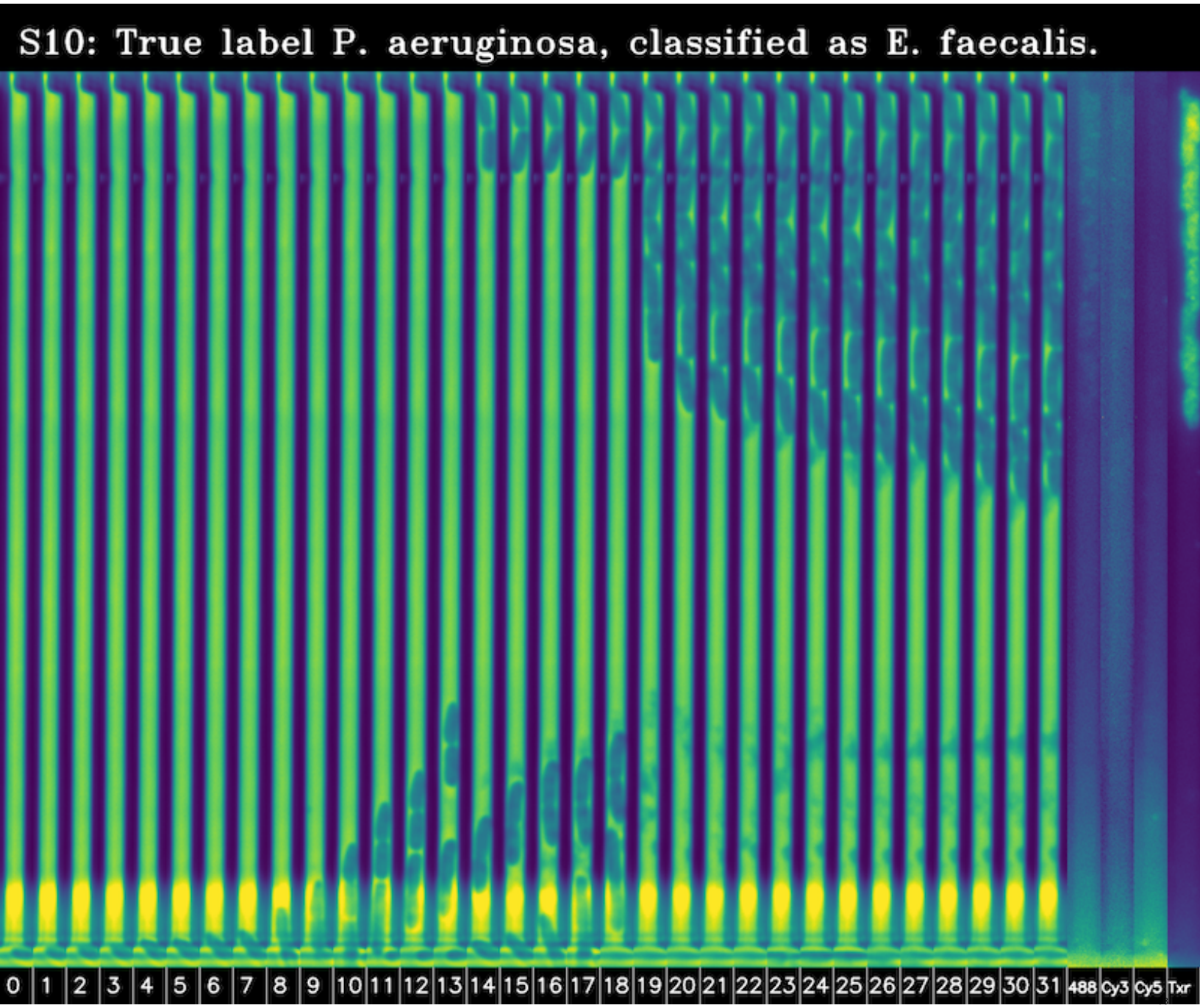
S7 Fig. Misclassified by ResNet-26 True label *P. aeruginosa*, classified as *E. coli*. Both species are rods with similar shapes and are easily confused. It appears to have been several species in the trap that avoided staining under heavy antibiotic treatment.



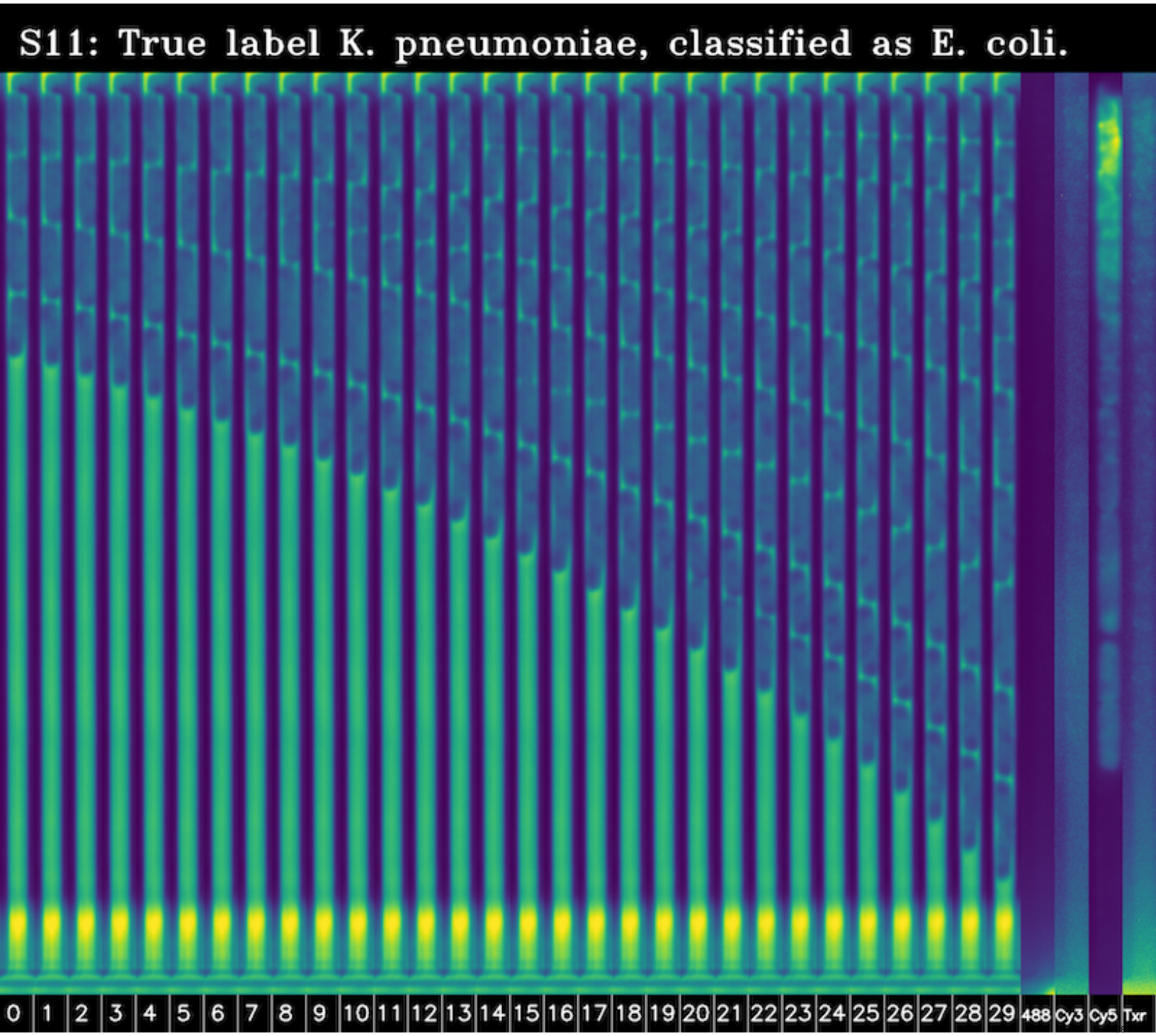
S8 Fig. Misclassified by ResNet-26 True label *K. pneumoniae*, classified as *E. coli*. Both species are rods with similar shapes and are easily confused.



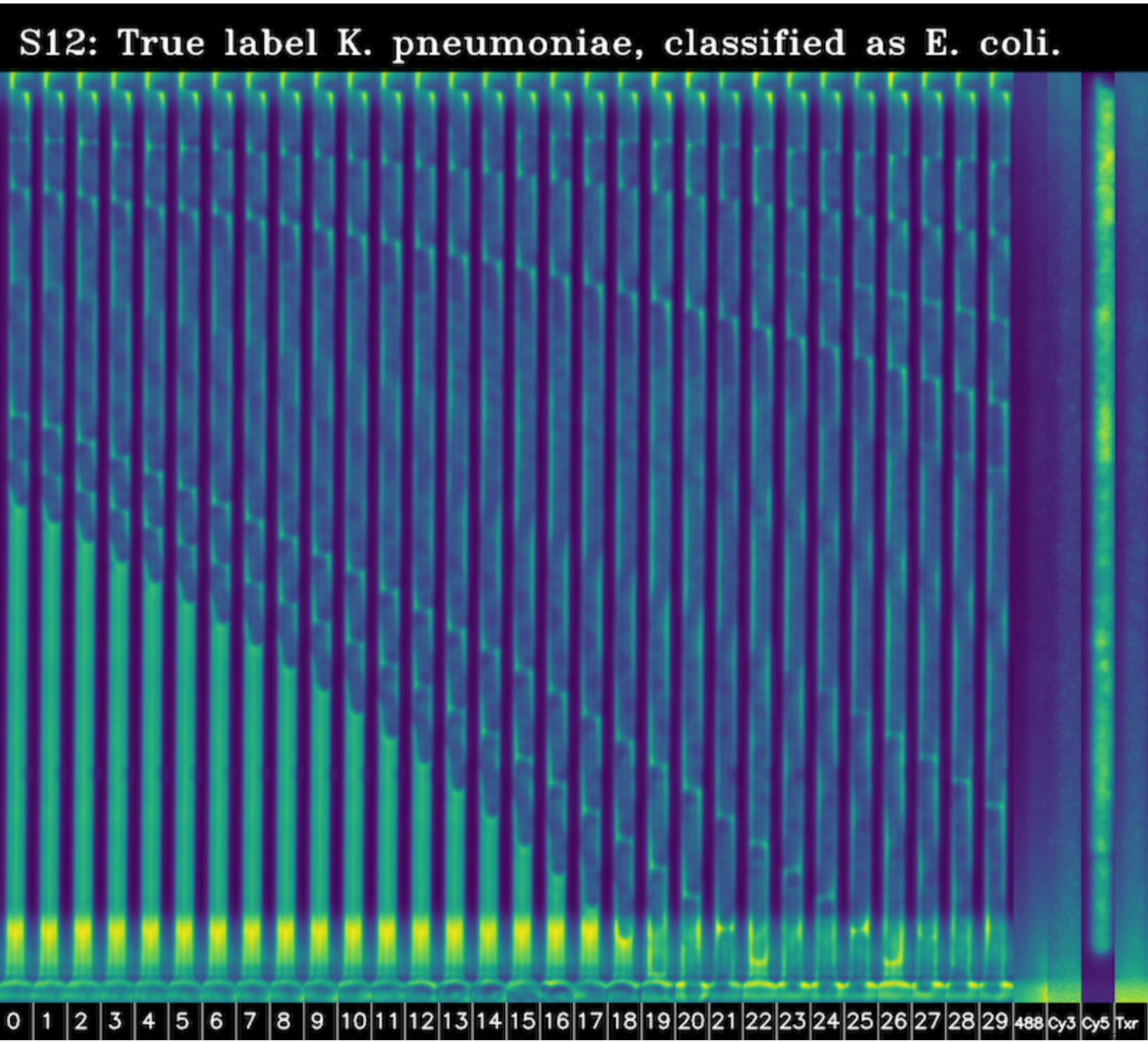
S9 Fig. Misclassified by ViT/B True label *E. coli*, classified as *K. pneumoniae*. Both species are rods with similar shapes and are easily confused.



S10 Fig. Misclassified by ViT/B True label *P. aeruginosa*, classified as *E. faecalis*. The ViT possibly confused the stop at the top of the trap as a coccus. The trap was empty in the first frame.



S11 Fig. Misclassified by ViT/B True label *K. pneumoniae*, classified as *E. coli*. Both species are rods with similar shapes and are easily confused.



S12 Fig. Misclassified by R(2+1)D True label *K. pneumoniae*, classified as *E. coli*. Both species are rods with similar shapes and are easily confused.

References

1. Riordon J, Sovilj D, Sanner S, Sinton D, Young EW. Deep learning with microfluidics for biotechnology. Trends in biotechnology. 2019;37(3):310–324.

2. Zhang Y, Jiang H, Ye T, Juhas M. Deep learning for imaging and detection of microorganisms. Trends in Microbiology. 2021;29(7):569–572.

3. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25.

4. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248–255.

5. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.

6. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. PMLR; 2015. p. 448–456.

7. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

8. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations; 2021.Available from: `https://openreview.net/forum?id=YicbFdNTTy`.

9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

10. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition; 2018. p. 6450–6459.

11. Tran D, Wang H, Torresani L, Feiszli M. Video classification with channel-separated convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 5552–5561.

12. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, et al. Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. p. 3202–3211.

13. Kandavalli V, Karempudi P, Larsson J, Elf J. Rapid antibiotic susceptibility testing and species identification for mixed samples. Nature Communications. 2022;13(1):6215.

14. Wang H, Ceylan Koydemir H, Qiu Y, Bai B, Zhang Y, Jin Y, et al. Early detection and classification of live bacteria using time-lapse coherent imaging and deep learning. Light: Science & Applications. 2020;9(1):118.

15. Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks. In: proceedings of the IEEE International Conference on Computer Vision; 2017. p. 5533–5541.

16. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700–4708.

17. Hay EA, Parthasarathy R. Performance of convolutional neural networks for identification of bacteria in 3D microscopy datasets. PLoS computational biology. 2018;14(12):e1006628.

18. Zieliński B, Plichta A, Misztal K, Spurek P, Brzychczy-Włoch M, Ochońska D. Deep learning approach to bacterial colony classification. PloS one. 2017;12(9):e0184554.

19. Mai DT, Ishibashi K. Small-Scale Depthwise Separable Convolutional Neural Networks for Bacteria Classification. Electronics. 2021;10(23):3005.

20. Smith KP, Kang AD, Kirby JE. Automated Interpretation of Blood Culture Gram Stains by Use of a Deep Convolutional Neural Network. Journal of Clinical Microbiology. 2017;56.

21. Panigrahi S, Murat D, Le Gall A, Martineau E, Goldlust K, Fiche JB, et al. Misic, a general deep learning-based method for the high-throughput cell segmentation of complex bacterial communities. Elife. 2021;10:e65151.

22. Touvron H, Vedaldi A, Douze M, Jégou H. Fixing the train-test resolution discrepancy. Advances in neural information processing systems. 2019;32.

23. Kandavalli V, Karempudi P, Larsson J, Elf J. Rapid antibiotic susceptibility testing and species identification for mixed samples. Nature Communications. 2022;13(1):6215.

24. Baltekin Ö, Boucharin A, Tano E, Andersson DI, Elf J. Antibiotic susceptibility testing in less than 30 min using direct single-cell imaging. Proceedings of the National Academy of Sciences. 2017;114(34):9170–9175.

25. Thevenaz P, Ruttimann UE, Unser M. A pyramid approach to subpixel registration based on intensity. IEEE transactions on image processing. 1998;7(1):27–41.

26. Hallström E, Kandavalli V, Ranefall P, Elf J, Wählby C. Phase contrast time-lapses of four bacteria species growing in microfluidic chips.; 2023. Available from: `https://doi.org/10.5281/zenodo.7888072`.

27. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. Unmasking Clever Hans predictors and assessing what machines really learn. Nature communications. 2019;10(1):1–8.

28. Carter B, Jain S, Mueller JW, Gifford D. Overinterpretation reveals image classification model pathologies. Advances in Neural Information Processing Systems. 2021;34:15395–15407.

29. Wightman R, Touvron H, Jegou H. ResNet strikes back: An improved training procedure in timm. In: NeurIPS 2021 Workshop on ImageNet: Past, Present, and

Future; 2021.Available from: https://openreview.net/forum?id=NG6MJnVl6M5.

30. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. PMLR; 2021. p. 10347–10357.

31. Mazzoli F. Lánczos interpolation explained; 2022. Available from: https://doi.org/10.5281/zenodo.7761401.

32. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, et al. The kinetics human action video dataset. arXiv preprint arXiv:170506950. 2017;.

33. Loshchilov I, Hutter F. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:160803983. 2016;.

34. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019;32.

35. Wightman R, Raw N, Soare A, Arora A, Ha C, Reich C, et al.. rwightman/pytorch-image-models: v0.8.6dev0 Release; 2023. Available from: https://doi.org/10.5281/zenodo.7527889.