ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

DeepMAP: Deep CNN Classifiers Applied to Optical Mapping for Fast and Precise Species-Level Metagenomic Analysis

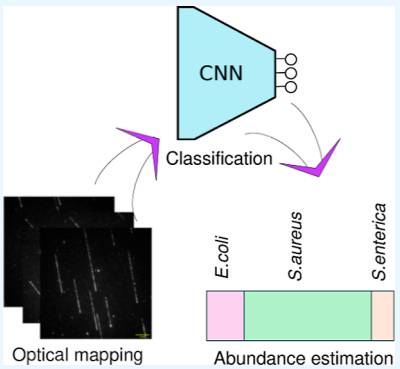
Sergey Abakumov, Elizabete Ruppeka-Rupeika, Xiong Chen, Arno Bouwens, Volker Leen, Peter Dedecker,* and Johan Hofkens

Cite This: *ACS Omega* 2025, 10, 9224−9232

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: DNA optical mapping is a powerful technique commonly used for structural variant calling and genome assembly verification. Despite being inherently high-throughput, the method has not yet been applied to highly complex settings such as species identification in microbiome analysis due to the lack of alignment algorithms that can both assign large numbers of reads in minutes and handle large database size. In this work, we present a novel genomic classification pipeline based on deep convolutional neural networks for optical mapping data (DeepMAP), which can perform fast and accurate assignment of individual optical maps to their respective genomes. We furthermore achieve a superior performance of DeepMAP in the presence of evolutionary divergent sequences, making it robust to the presence of unknown strains within metagenomic samples. We evaluate DeepMAP on genomic DNA extracted from bacterial mixtures, reaching species-level resolution with true positive rates of around 75% and a false positive rate of less than 1%, with measured classification speeds significantly outpacing those of previously developed approaches for high-density optical mapping data alignment.



INTRODUCTION

Recent advances in high-throughput sequencing have had a significant impact on the study of the microbiome, the collection of microorganisms that naturally inhabit the (human) body, including eukaryotes, prokaryotes, and viruses, and on how its composition relates to host well-being. Its composition within the human body has been implicated in a range of health-related processes, such as the effectiveness of immunotherapy in cancer treatment,[1] and is also heavily linked to obesity.[2] Currently, the two most common tools for microbiome profiling are 16s rRNA gene sequencing[3] and whole genome shotgun sequencing (WGS).[4] While WGS is able to identify the majority of organisms present within the sample, it remains an expensive procedure that requires long run times. 16s rRNA sequencing provides a faster and comparatively inexpensive approach but cannot be used if this RNA is not present, as is the case for viruses. Furthermore, the choice of primers[5] and amplification bias[6] may affect the outcome of this methodology.

DNA optical mapping is another emerging technology for metagenomic analysis. It relies on site-specific labeling of DNA molecules, typically by either restriction enzymes[7] or methyltransferases,[8] and analysis of the distribution of the labeled sites over the DNA fragments. The technique has been applied for structural variant detection,[9] improvement of de novo assembly,[10] and microbial and viral detection.[11] Structural variant (SV) calling and de novo assembly are typically performed only with DNA fragments ranging between 100 and 500 kbp in length that have a very low labeling density, which facilitates individual dye localization and alignment of fragment sizes using dynamic programming frameworks.[12] However, this methodology becomes poorly applicable in scenarios where the extracted DNA fragments are significantly shorter or in which dense labeling is required to distinguish many organisms or closely related variants.

The ability to efficiently analyze smaller and densely labeled DNA fragments is a crucial requirement for the application of DNA mapping to areas such as metagenomic analysis. However, efficient computational algorithms such as OM-Blast[13] or FANDOM[14] require that the absolute position of each individual label can be precisely determined, which in turn requires low labeling densities (no more than one label per kilobase) due to the limited spatial resolution of the imaging. We previously developed an analysis approach that could achieve recognition also for densely labeled samples based on calculating the cross-correlation between the experimental optical mapping trace and a reference sequence.[11] However, this analysis requires considerable runtime when applied to an increasing number of reference genomes and

Received: October 17, 2024
Revised: January 27, 2025
Accepted: February 6, 2025
Published: February 27, 2025

encounters difficulties when dealing with lower labeling efficiencies and/or variations of the DNA stretching along the fragments.

Applying high-density optical mapping to complex samples will require the development of more efficient analysis methodologies. We reasoned that the alignment of an experimental optical map to the reference genome is analogous to pattern classification, a task that is very efficiently performed by machine learning. Convolutional neural networks (CNNs) have been found to be particularly successful for such problems and are regularly applied to electroencephalography classification,[15] nanopore signal demultiplexing,[16] and the recognition of DNA binding motifs.[17] Furthermore, CNNs can be easily parallelized on GPU hardware, offering increased throughput.

While highly promising, deep learning has found limited use in the context of optical mapping analysis. Thus far, only large (150 kbp) DNA fragments could be classified with help of CNNs, where the CNN localizes individual emitter positions and therefore requires sparse labeling.[18] The state-of-the-art methods for high-density optical mapping data alignment furthermore require the presence of the exact strain sequence in the database. This, however, poses an issue for complex samples, where isolates might have diverged from their ancestor by acquiring mutations or developing structural variations, which causes exact alignment algorithms to perform poorly. CNNs may well be more robust against such variations and outperform known alignment algorithms given their proven ability to generalize across diverse training sets.

In this work, we investigated this approach by developing "DeepMAP", an analysis pipeline that classifies sequence-specifically labeled optical mapping data using CNN models. We evaluate DeepMAP on genomic DNA extracted from bacterial mixtures, reaching species-level resolution with true positive rates of around 75% and a false positive rate of less than 1%, with classification speeds significantly outpacing those of our previously developed approach for high-density optical mapping data alignment. The hierarchical nature of CNNs allows the algorithm to efficiently process information at both small and large scale, allowing it to operate on smaller fragments yet at the same time generalize to larger structural rearrangements via the intrinsic capability of the network to consider a broad range of length scales. Furthermore, CNNs are largely translation invariant, making them robust against the absence or presence of a particular feature and therefore mitigate artifacts introduced by deletions or inversions within the genome. We show an enhanced performance of DeepMAP in cases in which the exact strain is not present in the database, showing that DeepMAP can still identify it at the species level.

By providing a computationally more efficient analysis of the optical mapping data, as well as increased ability to handle sample diversity, DeepMAP promises to strongly expand the technology of optical mapping by allowing it to be applied to the unraveling of considerably more complicated samples.

METHODS

Bacterial Culturing. Three different bacterial cultures, *Escherichia coli*, *Salmonella enterica*, and *Vibrio harveyi*, were grown on LB agar (Invitrogen), at 37 °C overnight for *E. coli* and *S. enterica*, while *V. harveyi* was grown at 30 °C. Subsequently, they were inoculated in 5 mL of LB broth solution overnight. 0.5 mL of the cultured bacteria was first pelleted by centrifuging at 16,000*g* for 1 min and resuspended in 1 mL of distilled water. The high molecular weight (HMW) DNA was extracted using the Circulomics kit (Circulomics, Baltimore US) and the corresponding HMW protocol. Upon extraction, the DNA was left to homogenize overnight in a thermomixer at 50° and 300 rpm. The concentrations of each of the DNA stocks were measured on a Biodrop instrument before being stored at 4 °C until further use. Complex samples of bacterial DNA were prepared by mixing these stocks based on the desired genomic abundances of each of the species. In total, 3 mixtures were prepared: 1:1 (*E. coli*/*S. enterica*), 1:1 (*E. coli*/*V. harveyi*), and 1:1:1 (*E. coli*/*S. enterica*/*V. harveyi*).

M*Taq*I DNA Labeling and Purification. The DNA was labeled using the M*Taq*I enzyme (recognition sequence 5′-TCGA-3′) and the synthetic cofactor MTC-22.[19] The labeling mixture consisted of 1 μg of DNA and contained 1x CutSmart buffer (10x stock, NEB), a 50 μ M concentration of MTC-22, and a final concentration of 0.175 μ g/μ L of the M*Taq*I enzyme. In total, 40 μ L of the mixture was incubated at 60 °C for 1 h. Subsequently, 1.6 units of Proteinase K (0.8 u/μL, NEB) was added to the solution, with an additional 1 h incubation at 50 °C. The DNA was then purified using the protocol outlined in ref 11 0.02 g of agarose was dissolved in 1 mL of TAE buffer and melted at 70 °C. 27.2 μL of agarose solution was added to the labeling mixture and incubated for 30 min at 4 °C. Once the solution had gelated, it was left for 30 min in 1 mL of 1× TAE buffer (50×, Thermo Scientific). This wash step was repeated 4 times, and the buffer was replaced every time. The plug containing the DNA was then incubated for 15 min at 70 °C in order to melt the agarose and subsequently incubated for 45 min with 1 unit of agarose. After digestion, the droplet containing labeled DNA was left for 90 min to dialyze on top of the Millipore membrane (0.1 μ m, Merck). Once this step had been completed, 2 μ L of labeled DNA solution was diluted 5x and 1 μ L of 0.5 M MES buffer was added as to achieve 0.05 M concentration. The DNA was then subsequently stretched on the top of a glass slide, which had been covered with a Zeonex polymer using a procedure outlined in ref 11

Imaging. The samples were imaged using a Nikon Ti2 Eclipse inverted microscope equipped with a perfect focus system (PFS), allowing for automated focusing during scans. A rectangular area of 2 × 3 mm^2 containing between 500 and 600 images was scanned with the help of an automated stage. Image acquisition was performed using an oil immersion objective (NA 1.49, Nikon CFI SR HP Apochromat TIRF 100X Oil) and an additional 1.5x lens, with an exposure time of 0.4 s per image. The total area per image was 130 × 130 μm^2 with a virtual pixel size of 78.6 nm/pixel. The recognition sequence of *M.TaqI* occurs on average every 1 of 256 bp, resulting in an average distance of 0.15 μm between two adjacent dyes, which is lower than the diffraction limit of approximately 200 nm. Therefore, the resulting images contained continuous line signals produced by the optical maps. The images were subsequently segmented, by first thresholding the image, and the optical map was obtained in a similar manner to ref 11, as described in more detail in Supporting Information Note S1. Only DNA molecules that had a length of at least 42.3 kbp were segmented from the images.

Data Simulation. A simulation pipeline was implemented in order to generate large amounts of data for training the analysis models, including the effects of common errors encountered during optical mapping. An overview of the

https://doi.org/10.1021/acsomega.4c09485
ACS Omega 2025, 10, 9224−9232

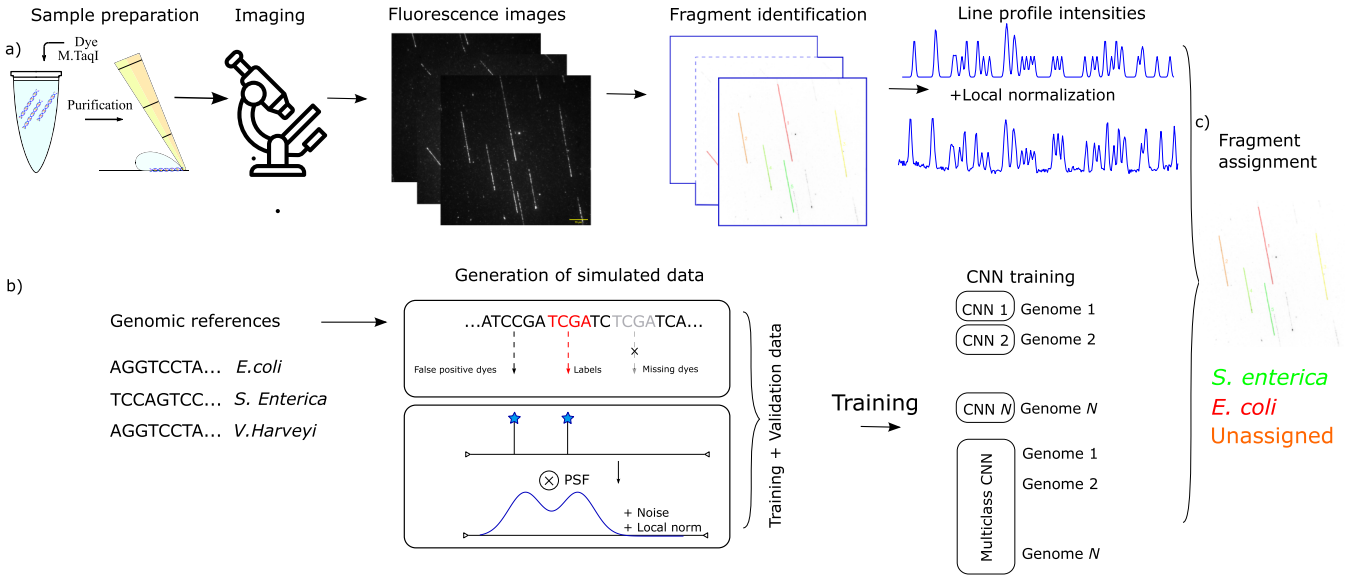


Figure 1. (a) Acquisition pipeline of the optical mapping procedure: extracted DNA is sequence-specifically labeled using the M.*Taq*I enzyme and is subsequently deposited on a Zeonex-covered substrate. The stretched DNA is imaged under a microscope, and individual DNA molecules are segmented in order to obtain line profiles of the individual DNA molecules. (b) Analysis methodology: genomic reference sequences are used to generate artificial line profiles for the corresponding organismal DNA, potentially including nonidealities. These profiles are then used to train and validate CNNs for classification. (c) Combining the experimental pipelines with the CNN-based classification allows individual DNA fragments to be assigned to particular species.

simulation pipeline is given in Figure 1. The overall simulation pipeline is discussed in Supporting Information Note S2. Briefly, reference genomes for the organisms used here were retrieved in the FASTA format from the NCBI database using the accession codes listed in Supporting Information Table S1. From these genomes, randomly positioned fragments were used to generate synthetic optical maps by converting the recognition sequences into a sparse array of dye positions. To simulate labeling imperfections, modifications such as variations in DNA stretching and labeling efficiency were added into the simulation pipeline. We additionally implemented a local randomization of the dye positions within a small window in order to provide the models with increased training diversity with the goal of matching closely related species such as different strains of a given organism. Although the simulations captured the majority of variation present in the optical mapping pipeline, gauging some of the experimental variations proved to be complicated, and therefore, these were not included in the simulation pipeline. These include errors stemming from improper focusing of the PFS during scanning, variation of the PSF across the field, and nonlinear variation of the stretch factor.

In addition to the optical maps calculated from actual genomic sequences, we also generated optical maps from entirely randomly generated sequences with label densities ranging from 1 to 6.8 labels/kbp, in line with the site densities typically found in bacterial genomic sequences. These were then used in training as a model for optical maps not arising from one of the reference species.

Data generation for the training and validation data sets (95/5 split) was handled separately as to avoid any overlap between both. The amount of training data has an effect on the generalization performance of the network. The optimal amount of sampling required for training our classifier is further discussed in Supporting Information Note S3. Each of the networks was trained using an adaptive moment estimation (ADAM) optimizer (learning rate: 10^{-3}) with a batch size of 256 traces. Binary cross entropy was used as the loss function. The loss for training and validation was stored upon the end of each epoch. To obtain an optimal set of weights, early stopping was used, and the set of parameters corresponding to the epoch with the lowest validation loss was selected.

RESULTS AND DISCUSSION

A conceptual overview of the proposed analysis methodology is shown in Figure 1, consisting of the acquisition of experimental optical maps (Figure 1a) that are then analyzed by using DeepMAP (Figure 1b). The experimental data generation strategy has been described in detail in the Supporting Information Note S2. Briefly, DNA is first sequence-specifically labeled with fluorescent labels. The labeled fragments are then stretched on a cover glass and visualized using fluorescence microscopy. The resulting images are analyzed in order to obtain line profiles showing the observed fluorescence intensity along each fragment. The task of DeepMAP is to compare these profiles with synthetic profiles generated from genomic sequences of particular organisms in order to determine whether DNA from these organisms is present within the sample and, if so, what their abundance is. To do so, we apply CNNs that have been trained on simulated line profiles calculated from the genomic sequences of the organisms in question in order to recognize corresponding DNA fragments within the optical mapping data.

A key requirement for CNN-mediated analysis is the availability of a large amount of data for training. While it is theoretically possible to use experimental data sets, acquiring a large number of optical maps is time- and resource-intensive and may be difficult to achieve on species that cannot be cultured directly.[23] As a result, we started by developing a simulation framework that allows the generation of simulated optical mapping fragments given an input genome sequence,

https://doi.org/10.1021/acsomega.4c09485
ACS Omega 2025, 10, 9224−9232

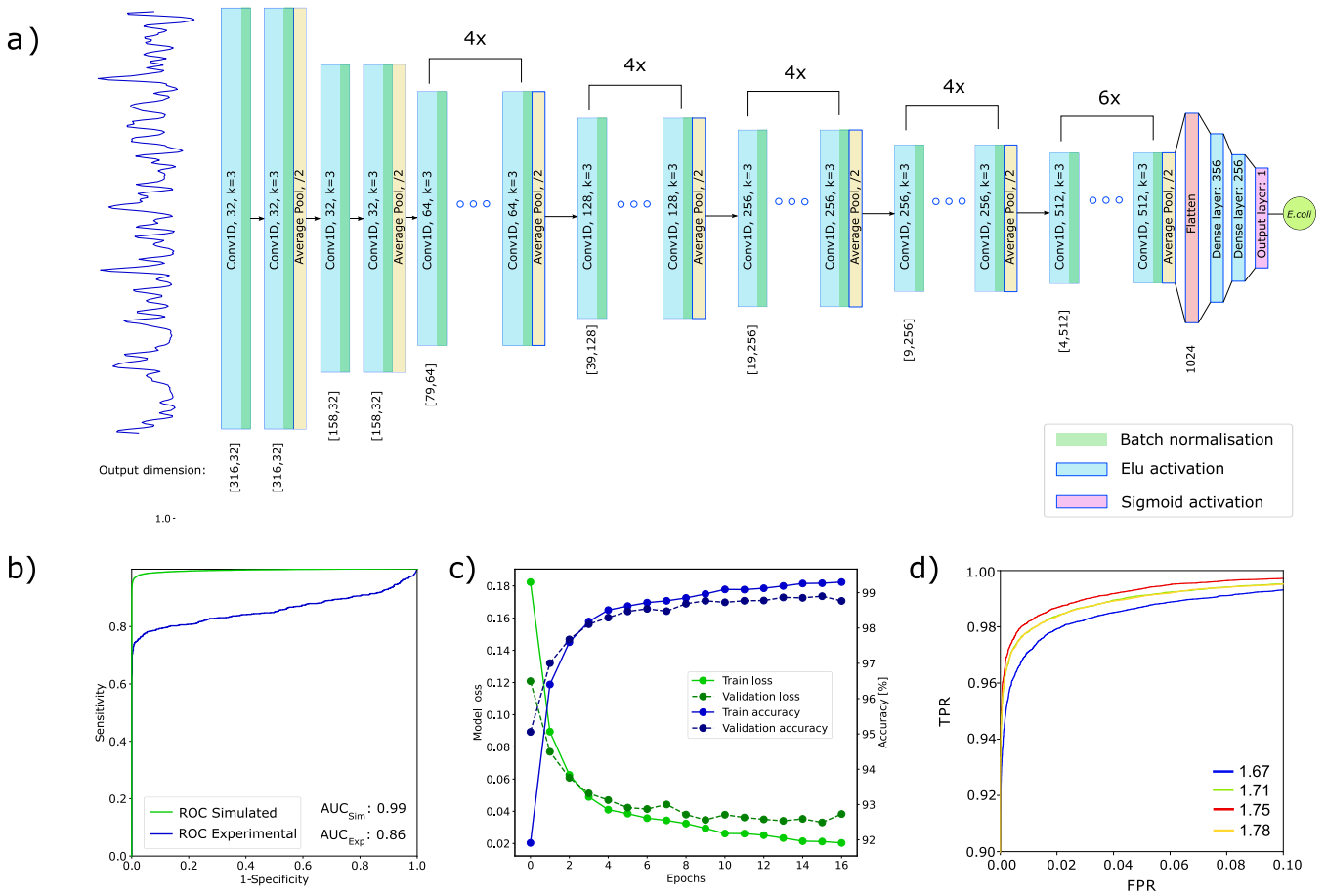


Figure 2. (a) Outline of the architecture for the single CNN, depicting the kernel sizes (k), the number of filters, and the topology. (b) ROC curve for the experimental (*E. coli*) and simulated data, with the control being a data set from *S. enterica*. (c) Training curves for the network: accuracy and model loss for validation and training data. (d) Effects of variations of the stretch factor on the ROC based on the simulated data.

augmented to account for common experimental imperfections such as missing labels or mislabeling. These data are then used for both training and validation of the network.

The optical maps generated with our experimental approach have a high degree of local correlations, particularly due to diffraction limited imaging by the microscope, while in the long-range, these are mostly uncorrelated. Accordingly, we selected a CNN for this work since a CNN encoder should be efficient in compressing the data prior to classification. The encoder part of the CNN architecture was based on architectures that were shown to work well in previous work based on pattern recognition and DNA sequence classification.[16] We compared 4 different architectures, as is detailed in Supporting Information, ultimately selecting the network depicted in Figure 2a. This network consists out of 26 1D-convolutional layers with a fixed kernel size of 3 and exponential linear unit (ELU) activations. The commonly used rectified linear unit (ReLU) activation was not used in our approach in order to circumvent the dying ReLU problem encountered in the deep CNN.[24]

The training of each network was performed as described in the Methods section of this work. We initially focused on a binary classification approach, where the neural network is provided with a line profile from an optical map and asked to classify whether this map originates from the reference genome on which the network was trained. By thresholding the final score, we can obtain the class of the fragment. If it indeed originated from the reference genome and is correctly classified by the CNN, then it is regarded as a true positive (TP). The respective true positive rate (TPR), or the sensitivity, is defined as the ratio of the number of fragments from the positive class correctly assigned to the positive class to the total number of positive class fragments in the test data set. If the CNN recognized a map from another organism, then this is regarded as a false positive (FP). The respective false positive rate (FPR), or 1-specificity, is then defined as the rate of falsely identified fragments from a negative class as a positive class to the total number of negative class fragments in the test data set. If the score of the optical map does not meet a particular threshold, then the optical map is considered to be unassigned. This threshold was chosen based on the receiver operator characteristics curve (ROC), constructed from experimental validation data in Figure 2b, so as to maintain a false positive rate of less than 1%. The network model used in this work is shown in Figure 2a with its training loss depicted in part 2c. As expected, the network becomes more accurate for each training epoch, though overfitting becomes apparent as well. In order to minimize this, we have trained the network for a fixed number of 20 epochs as during experimentation, the network started to overfit beyond this number. We selected the set of weights with the lowest validation loss from all epochs (Supporting Information Table S4). The ROCs for experimental (*E. coli* data set) and simulated data sets are plotted in Figure 2b, with the corresponding experimental and simulated

https://doi.org/10.1021/acsomega.4c09485
ACS Omega 2025, 10, 9224−9232

S. enterica controls. The sensitivity for simulated data reaches 98%, while the experimental data show a sensitivity of around 75%, similar to cross-correlation. Both show an excellent false positive rate, indicating a highly specific classification performance.

Our CNN architecture required us to specify a fixed size of the DNA fragment length. The majority of extraction protocols are able to yield fragments ranging from 20 to 60 kbp, although longer lengths typically require careful sample handling and often complicated extraction procedures. Shorter fragment lengths, on the other hand, lead to reduced experimental requirements but are likely to reduce the sensitivity and specificity. We explored this effect by training various CNNs on simulated optical maps calculated from the *E. coli* genome, supplemented with simulated maps from *S. enterica* that allowed us to estimate the false positive rate. An overview of these results is given in Table 1. We chose a fragment length of 42.3 kbp as a compromise between an acceptable false positive rate and what can realistically be obtained with state-of-the-art high-molecular-weight DNA extraction methods. Since the experimental segmented data contained optical maps that exceed the input length of 42.3 kbp, we have cropped these to the required input length. However, if the cropped map is classified to a genome g_i by the CNN, the entire length of the uncropped optical map (in kbp) is assigned to the genome g_i so as to avoid data loss.

Table 1. Effect of the Input Length on the Sensitivity and False Positive Rates at a Fixed Threshold, Evaluated on Experimentally Acquired Optical Mapping Data from *E. coli* MG1655 and a False Positive Control Constituting Experimental Optical Maps Originating from *S. enterica*

input length	model loss	sensitivity (%)	FPR (%)
34.2 kbp	0.1075	66.9	3.6
42.3 kbp	0.0435	71.9	1.7
47.7 kbp	0.0187	77.6	1.1
55.1 kbp	0.0096	78.0	0.3

We also explored the influence of DNA stretching on the classification, reflecting variations in the local flow of the droplet during deposition on a Zeonex-covered slide. Typical stretch factors range from 1.68 to 1.75.[11] We found that the trained CNN is able to perform well on a variety of stretch factors even when trained on a single stretch factor; thus, only a single stretch factor of 1.75 was chosen for training and validation, minimizing the amount of data and training times required. The ROCs for various simulated stretch factors are plotted in Figure 2d.

We then compared the previously developed cross-correlation approach[11] with the CNN-based methods introduced in this work by applying these to experimental data sets containing DNA from a single bacterial species. To deliver acceptable performance, we optimized the cross-correlation approach so that it could deal with bacterial genomes, which are much larger than the viral genomes to which this method was previously applied to. The changes made as part of this optimization are detailed in Supporting Information Note S5. The comparison results for *E. coli* and *V. harveyi* are plotted in Figure 3a,b. The sensitivity for the CNN classifier is given in Figure 3d. The genomic abundances for both chromosomes of *V. harveyi* are also plotted in Figure 3c and correspond well to the reference abundance of the chromosomes. During this analysis, we have omitted the additional plasmid in the genome of *V. harveyi* due to its small size (0.09 Mbp) in comparison to the chromosomes (3.77 and 2.20 Mb). We have calculated the relative genomic abundance of genome g_i as follows

$$g_i = \frac{\sum \text{lengths of fragments assigned to species } i}{\sum \text{lengths of fragments assigned to all species}} \quad (1)$$

the ratio between the sum of lengths of optical maps assigned to that genome and the total length sum of all optical maps assigned to all genomes. Fragments that were assigned to the random category were considered as unassigned and therefore removed from the relative abundance calculation.

Although both approaches perform well on experimental data and are able to recover the relative abundances, the sensitivity of the proposed CNN method is lower than that achieved with cross-correlation (Figure 3b and Table 2). The sensitivities of the networks also appeared to differ slightly when applied to optical maps from different bacterial species (Figure 3d), indicating that some reference genomes lead to optical maps that are better suited to recognition by these CNNs, potentially introducing bias. Compared to cross-correlation, however, the CNN-based approach is much more computationally efficient, outperforming the cross-correlation approach by a factor of 10^3 in classification speed (see Table 2). Given an input length of 42 kbp and a measured classification speed of 0.6 ms per single fragment (evaluated on a single Nvidia Quadro P4000 GPU), the overall time to classify 1 Gb of genomic material is around 15 s for the single CNN classifier. For a modest database of ∼1000 genomes, this yields a classification time of ∼4.2 h.

We wondered to what extent the CNN-based analysis could recognize different organismal strains in cases in which the exact strain is not present in the database. We simulated such a case in Figure 4, applying the *E. coli* K-12 MG1655-trained CNN to the analysis of *E. coli* O157H7 str. Sakai and O103H3 str 12009, with which MG1655 shares 97.86 and 98.37% average nucleotide identity (ANI), respectively. We also compared with the result of cross-correlation with K-12 MG1655. We have constructed the ROCs by sampling optical maps uniformly from the genome of each strain, which were treated as true positive class. The false positive class included optical maps that are randomly generated with label densities ranging from 1 to 6.8 labels/kbp. If an optical map from an *E. coli* strain was classified by the CNN as *E. coli*, this map was considered as part of true positive class. However, if a randomly generated optical map was classified as *E. coli*, this map was considered part of the false positive class. The ROCs for this analysis show an enhanced performance of CNNs in recognizing such strains compared to that of cross-correlation, reaching up to a 2× increase in TP rate, while maintaining the same FP rate (see Figure 4a). To realize this performance, we added additional shuffling during the generation of the training data, as described in the Methods section. To visualize this in more detail, we plotted the matched positions within the genome for both classification approaches and compared them to full genomic alignment computed with Mauve[25] of str. O157H7 to K12 str MG1655 (Figure 4b). The improved performance of DeepMAP is immediately clear, especially for regions that are broken up by insertions. These can still be matched with our DeepMAP approach, while the cross-correlation is not able to obtain proper alignments. This shows

https://doi.org/10.1021/acsomega.4c09485
ACS Omega 2025, 10, 9224−9232

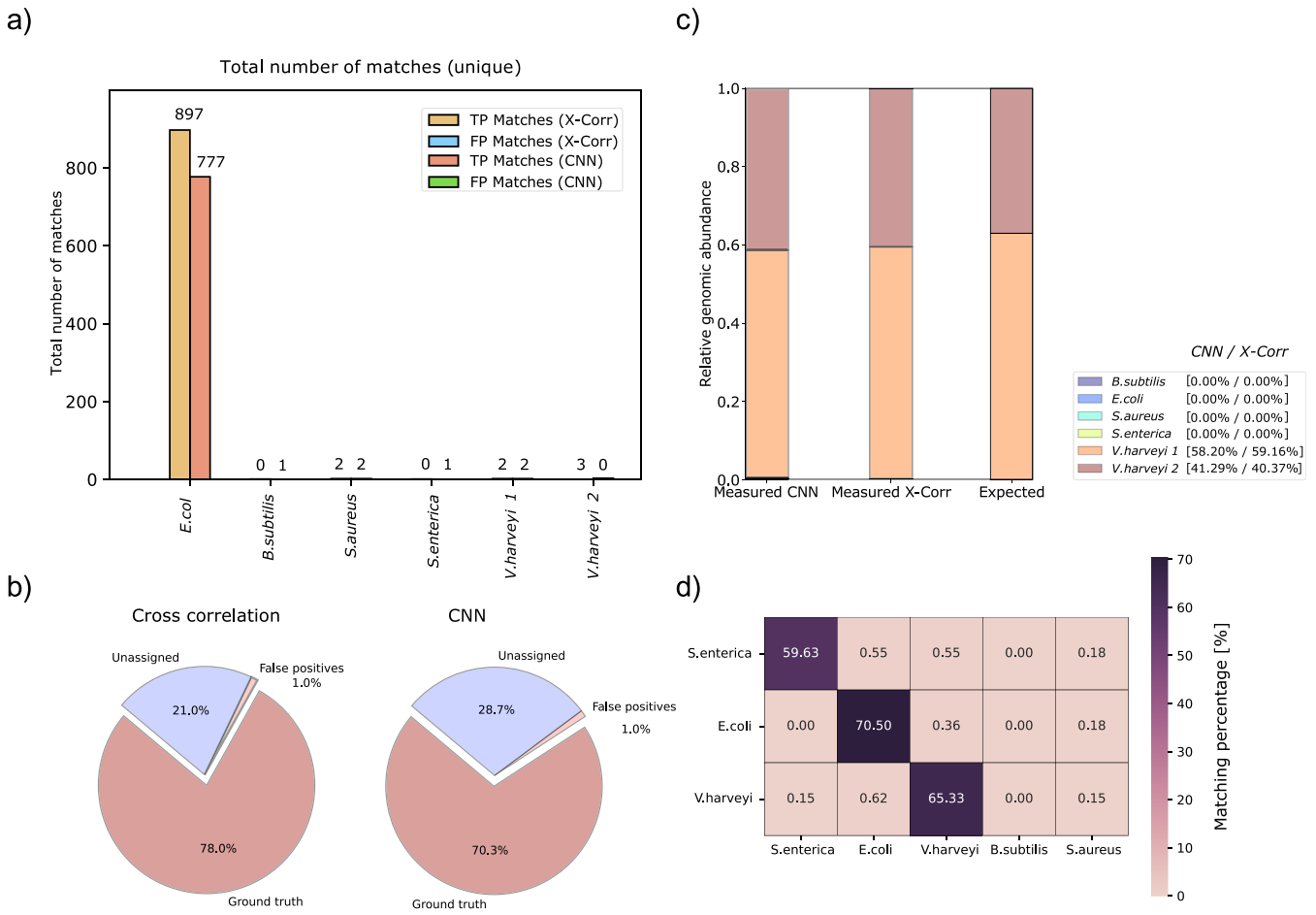


Figure 3. (a) Absolute number of matches to the *E. coli* reference, including 2 false positive genomes. (b) Sensitivity of a single-class classifier: ground truth is detected *E. coli*, unassigned did not meet the threshold, and false positives are detected by other networks. (c) Relative genomic abundance of the two *V. harveyi* chromosomes. (d) Matching matrix for the three single species datasets, with two *V. harveyi* chromosomes pulled together in order to determine the overall sensitivity.

Table 2. Summary of the Performances of the CNN Ensembles in Absolute Numbers of Classified Traces for Each of Single Species Datasets[a]

data set	total maps	true positive		false positive		time per map (s)	
		CNN	X-Corr	CNN	X-Corr	CNN	X-Corr
E. coli	1111	777	897	6	7	0.0006	0.93
S. enterica	544	336	406	9	10	0.0006	0.87
V. harvey	648	408	424	7	7	0.0006	0.86

[a]The number of maps differs between various genomes due to the variations in deposition density. We further pooled the analysis from both CNNs for two chromosomes of *V. harveyi* together. Measured times per single optical map are also given, showing a better performance of CNNs than that of the cross-correlation approach.

that the CNN-based approach can recognize divergent genome sequences at the species level.

We next investigated the possibility of applying the CNNs to the analysis of experimental datasets containing known-abundance mixtures of genomic DNA from various species. We created multispecies samples by simply mixing the precalculated amounts of extracted genomic DNA from corresponding bacteria. In principle, the analysis of such mixtures can take place in two ways: an ensemble containing multiple CNNs trained for individual species can be applied to the same dataset, with each CNN determining which of the DNA fragments originates from "their" species and which ones cannot be matched, assigning the fragment to whichever of these "single-class" CNNs calculated the highest score. A second strategy is to train a single CNN that can distinguish between multiple different species, known as a multiclass classifier.

We initially investigated the use of multiple single-class classifiers. Figure 5a–c shows the analysis results for three different mixtures as well as the analysis results obtained using the previously published methodology based on cross-correlation. The composition determined with CNN ensembles shows a deviation from the expected DNA content. Despite this deviation from the expected abundances, the CNN results are in agreement with the results obtained with cross-correlation, showing a near one-on-one correspondence. We conclude that while the current analysis strategies do not fully reproduce the expected species distribution in the sample, the CNN-based methodology matches the performance of the cross-correlation-based approach while being several orders of magnitude faster.

The performance of the multiclass classifiers is further investigated in Figure 6, starting with the training and validation losses in Figure 6a. We performed a separate evaluation on the *V. harveyi* and bacterial mixtures and compared the results with the cross-correlation results. We also

https://doi.org/10.1021/acsomega.4c09485
ACS Omega 2025, 10, 9224−9232

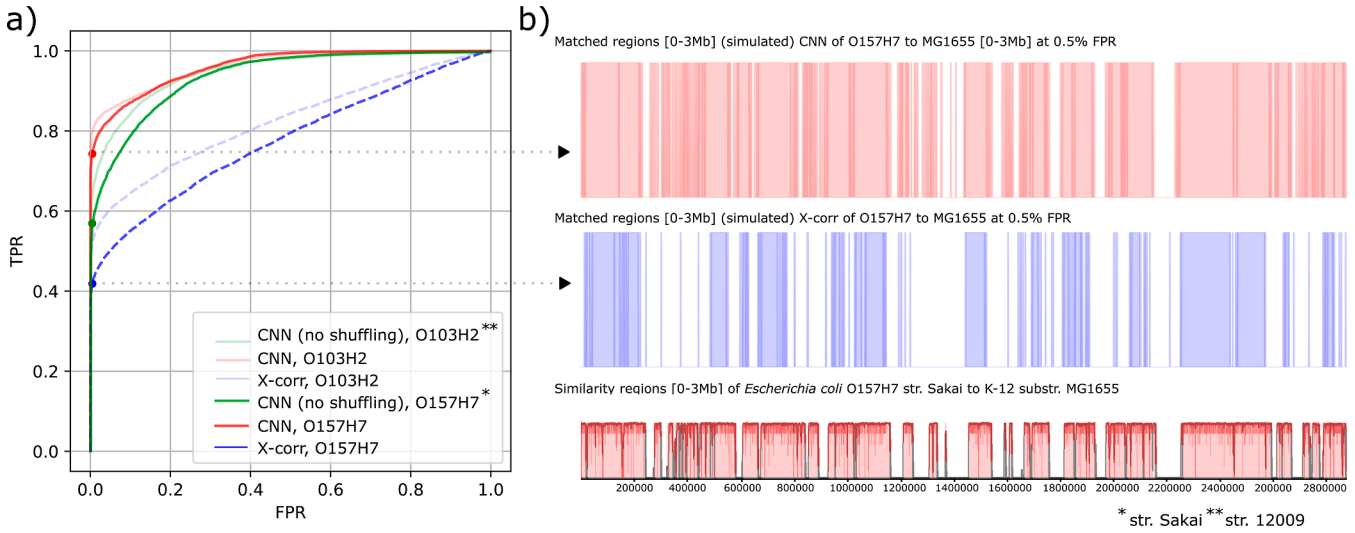


Figure 4. (a) ROC of CNNs and X-corr on simulated data from two related strains of *E. coli*. The CNNs show an enhanced performance compared to cross-correlation if the exact genome is not present in the database. (b) Aligned regions of CNN and X-corr alignments on O157H7 str. Sakai corresponding to 0.5% FPs. The positions within the genome that could be matched are depicted in red for DeepMAP and in blue for cross-correlation. The bottom graph corresponds to the MAUVE alignment of the O157H7 to MG1655 strain.

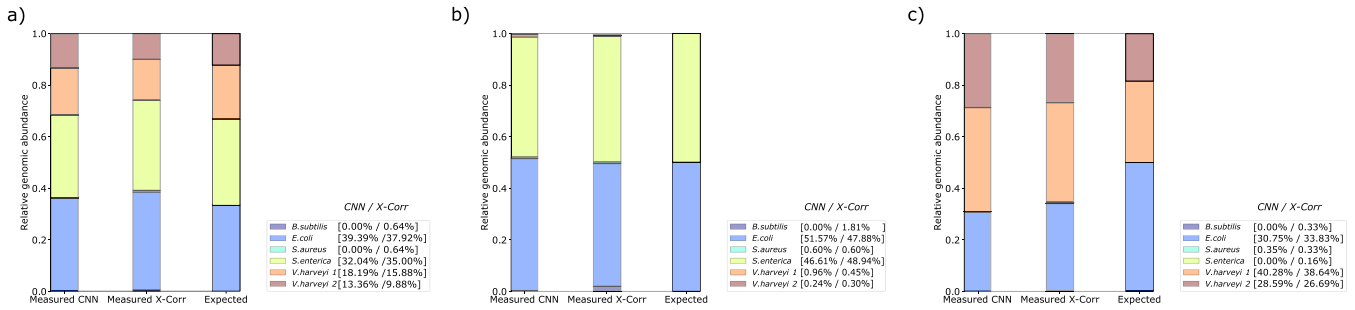


Figure 5. (a–c) Relative genomic abundances determined from the predicted approximate DNA content in the mixture based on the DNA stock concentrations (expected values) determined by CNN classifiers. *S. aureus* and *B. subtilis* are included as controls and are not present in either of the mixtures.

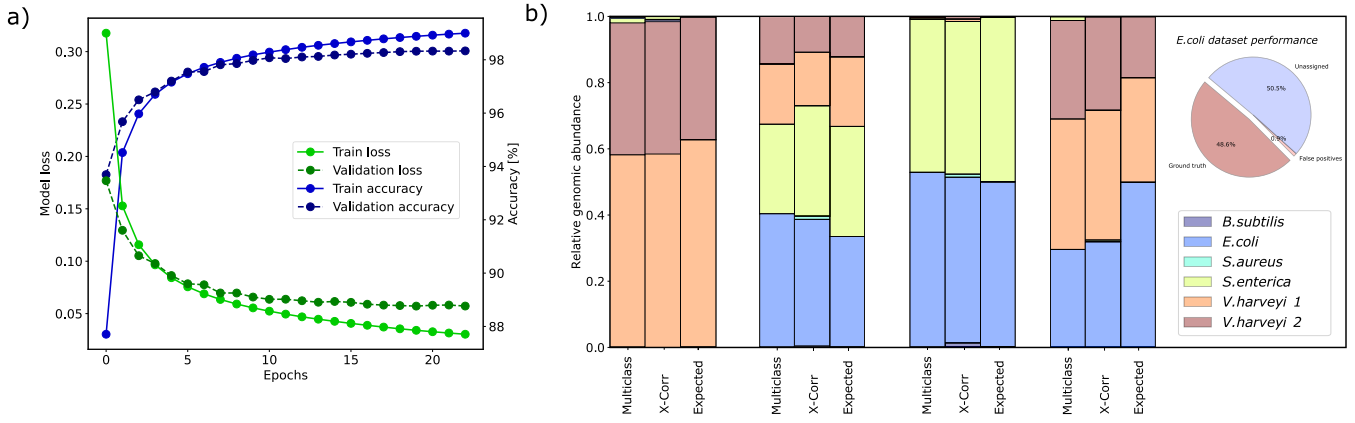


Figure 6. (a) Loss and accuracy for training the multiclass classifier. (b) Performance of multiclass classifiers on aforementioned data sets and their comparison to the reference cross-correlation results and their respective expected values. The results between the two methodologies compare fairly, and the offset can likely be explained by the difference in sensitivity.

report the classification metrics on the *E. coli* dataset, which served as the benchmark for the single genome CNN classifier. These results are plotted in Figure 6b and show a reasonable correspondence with the cross-correlation results for 3 out of 4 datasets. The predicted relative abundances are nearly identical, differing by a few percent from those predicted by cross-correlation. The offset from the real target abundances can be similarly explained by experimental error. The CNN ensembles yield similar results, as can be inferred from Figure 5. However, as can be seen from the pie chart, the sensitivity drops by ≈20% with respect to the single genome CNN approach, making the multiclass classifier less sensitive in

https://doi.org/10.1021/acsomega.4c09485
ACS Omega 2025, 10, 9224−9232

comparison. With the current implementation, an ensemble of single-class specifiers yields the highest classification accuracy when compared to a multiclass implementation. However, multiclass architectures can in principle handle the classification of more than four species simultaneously, though this would require an extensive architecture search to determine the optimal hyperparameters of the network since training the network on classifying more information becomes increasingly difficult with an increased number of genomes.

CONCLUSIONS

In this work, we have introduced deep CNN classifiers as a strategy for the classification of maps produced from optical mapping experiments. We find that this approach delivers a classification sensitivity similar to that of an established approach relying on cross-correlation but delivers a speed-up of several orders of magnitude.

The current implementation can take advantage of this efficiency to deliver good scaling with respect to the size of the reference database, yielding a total of 4.2 h classification times for 1 gigabase of data (approximately 24,000 DNA fragments) with a database size of 10^3 genomes. Multiclass classifiers can help further improve this performance, though we here found that this led to a small increase in the false positive rate. This could be compensated by adjusting the threshold on the final score provided by the CNN, at the cost of the sensitivity, or potentially revising the training method to take advantage of transfer learning.[20,21]

In principle, the training data for the CNNs can be provided from actual experiments performed on single-species samples or using simulated data generated by an appropriate simulation framework. The use of simulations offers the advantage of fast and convenient data generation and can also generate data when culturing of the required species becomes difficult or infeasible. Our results show that the use of such simulations readily results in good to very good performance by using comparatively modest effort. The utilization of random fragments as data for the second class could potentially be replaced by more rigorous open-set classification approaches, such as ones based on class modeling[22] or open-set CNN classifiers.[26]

A further advantage of CNNs is that this approach also provides comparatively straightforward adaptation to non-idealities such as lower labeling efficiencies, mislabeling, or heterogeneity in the DNA stretching, by incorporating such features in the data used to train the network. This flexibility is much more difficult to implement in classical algorithms. This is partially due to the increased depth of our convolutional encoder and average pooling layers before the classifier. The increased depth of our CNN model aids in better feature extraction from the optical maps, while the average pooling layers allow for increased information retention, which results in an enhanced classification performance and better generalization to experimental errors. We took further advantage of this ability by training the network to recognize evolutionarily divergent sequences. One disadvantage of the CNN-based approach, however, is the requirement for fragments to have a fixed length, here, 42 kbp, which requires shorter fragments to be discarded and longer fragments to be artificially cut. A possible way to circumvent these issues would be to utilize neural network architectures that are insensitive to variations in length input, such as recurrent neural networks; however, this may come at a cost of classification performance.

In conclusion, we have presented DeepMAP, a novel genomic classification pipeline based on deep CNNs for optical mapping data, which is able to perform fast and accurate assignment of individual optical maps to their respective genomes. We expect that DeepMAP can enable the application of optical mapping to much more complex samples, by offering much faster processing as well as the ability to adapt to the particulars and perhaps nonidealities present in the sample or preparation procedures.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c09485.

Image segmentation; example of a scanned image and its segmentation; simulation framework; description of reference genomes used in the analysis; illustration of potential contribution of the background to the DNA map; simulation pipeline for obtaining data; effect of the amount of training data; overview of the performance metrics for the models; choice of model architecture; CNN architectures covered; execution times, losses, and sensitivities for each of the model; adjusted cross-correlation-based analysis; distribution of cross-correlation scores from a single alignment; and comparison of ROCs Supporting Information.docx: supplementary file accompanying the manuscript The DeepMAP source code can be accessed from https://github.com/SAbakumov/DeepMAP (PDF)

AUTHOR INFORMATION

Corresponding Author

Peter Dedecker − *Department of Chemistry, Laboratory for Nanobiology, KU Leuven, 3000 Leuven, Belgium;* orcid.org/0000-0002-1882-2075; Email: peter.dedecker@kuleuven.be

Authors

Sergey Abakumov − *Department of Chemistry, Laboratory for Nanobiology, KU Leuven, 3000 Leuven, Belgium;* orcid.org/0000-0002-3834-5966

Elizabete Ruppeka-Rupeika − *Department of Chemistry, Laboratory for Molecular Imaging and Photonics, KU Leuven, 3000 Leuven, Belgium;* orcid.org/0000-0002-7093-8147

Xiong Chen − *Department of Chemistry, Laboratory for Molecular Imaging and Photonics, KU Leuven, 3000 Leuven, Belgium*

Arno Bouwens − *Perseus Biomics, 3300 Tienen, Belgium*

Volker Leen − *Perseus Biomics, 3300 Tienen, Belgium*

Johan Hofkens − *Department of Chemistry, Laboratory for Molecular Imaging and Photonics, KU Leuven, 3000 Leuven, Belgium;* orcid.org/0000-0002-9101-0567

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c09485

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by funds from the VLAIO mandate KRIS:0508000002079 to S.A. and from the European Research

https://doi.org/10.1021/acsomega.4c09485
ACS Omega 2025, 10, 9224−9232

Council through grant 714688 NanoCellActivity. J.H. acknowledges financial support from the Research Foundation-Flanders (FWO, grant no. G0C1821N and ZW15_09-G0H6316N), from the Flemish government through long-term structural funding Methusalem (CASAS2, Meth/15/04), and from the MPI as MPI fellow.

REFERENCES

(1) Dai, Z.; Zhang, J.; Wu, Q.; Fang, H.; Shi, C.; Li, Z.; Lin, C.; Tang, D.; Wang, D. Intestinal microbiota: a new force in cancer immunotherapy. *Cell Commun. Signal.* **2020**, *18*, 90.
(2) Muscogiuri, G.; Cantone, E.; Cassarano, S.; Tuccinardi, D.; Barrea, L.; Savastano, S.; Colao, A. Gut microbiota: a new path to treat obesity. *Int. J. Obes. Suppl.* **2019**, *9*, 10−19.
(3) Kuczynski, J.; Lauber, C. L.; Walters, W. A.; Parfrey, L. W.; Clemente, J. C.; Gevers, D.; Knight, R. Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* **2012**, *13*, 47−58.
(4) Ranjan, R.; Rani, A.; Ahmed, M.; McGee, H. S.; David, L. P. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* **2016**, *469*, 967−977.
(5) Tremblay, J.; Singh, K.; Fern, A.; Kirton, E. S.; He, S.; Woyke, T.; Lee, J.; Chen, F.; Dangl, J. L.; Tringe, S. G. Primer and platform effects on 16S rRNA tag sequencing. *Front. Microbiol.* **2015**, *6*, 771.
(6) Brooks, J. P.; Edwards, D. J.; Harwich, M. D.; Rivera, M. C.; Fettweis, J. M.; Serrano, M. G.; Reris, R. A.; Sheth, N. U.; Huang, B.; Girerd, P.; Strauss, J. F.; Jefferson, K. K.; Buck, G. A. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* **2015**, *15*, 66.
(7) Chan, S.; Lam, E.; Saghbini, M.; Bocklandt, S.; Hastie, A.; Cao, H.; Holmlin, E.; Borodkin, M. Structural Variation Detection and Analysis Using Bionano Optical Mapping. In *Copy Number Variants: Methods and Protocols*; Springer: New York, 2018; pp 193−203. −.
(8) Neely, R. K.; Dedecker, P.; Hotta, J.; Urbanavičiūtė, G.; Klimašauskas, S.; Hofkens, J. DNA fluorocode: A single molecule, optical map of DNA with nanometre resolution. *Chem. Sci.* **2010**, *1*, 453−460.
(9) Soto, D. C.; Shew, C.; Mastoras, M.; Schmidt, J. M.; Sahasrabudhe, R.; Kaya, G.; Andrés, A. M.; Dennis, M. Y. Identification of Structural Variation in Chimpanzees Using Optical Mapping and Nanopore Sequencing. *Genes* **2020**, *11*, 276.
(10) Howe, K.; Wood, J. M. D. Using optical mapping data for the improvement of vertebrate genome assemblies. *GigaScience* **2015**, *4*, 10.
(11) Bouwens, A.; Deen, J.; Vitale, R.; D'Huys, L.; Goyvaerts, V.; Descloux, A.; Borrenberghs, D.; Grussmayer, K.; Lukes, T.; Camacho, R.; Su, J.; Ruckebusch, C.; Lasser, T.; Van De Ville, D.; Hofkens, J.; Radenovic, A.; Frans Janssen, K. P.; Kris, P. Identifying microbial species by single-molecule DNA optical mapping and resampling statistics. *NAR:Genomics Bioinf.* **2019**, *2*, lqz007.
(12) Nagarajan, N.; Read, T. D.; Pop, M. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics* **2008**, *24*, 1229−1235.
(13) Leung, A. K. Y.; Kwok, T.-P.; Wan, R.; Xiao, M.; Kwok, P.-Y.; Yip, K. Y.; Chan, T.-F. OMBlast: alignment tool for optical mapping using a seed-and-extend approach. *Bioinformatics* **2017**, *33*, 311−319.
(14) Dehkordi Siavash, R.; Luebeck, J.; Bafna, V. FaNDOM: Fast nested distance-based seeding of optical maps. *Patterns* **2021**, *2*, 100248.
(15) Craik, A.; He, Y.; Contreras-Vidal, J. L. Deep learning for electroencephalogram (EEG) classification tasks: a review. *J. Neural. Eng.* **2019**, *16*, 031001.
(16) Ryan, R. W.; Judd, L. M.; Holt, K. E. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLoS Comput. Biol.* **2018**, *14*, No. e1006583.
(17) Zhou, J.; Lu, Q.; Xu, R.; Lin, G.; Wang, H. CNNsite: Prediction of DNA-binding residues in proteins using Convolutional Neural Network with sequence features *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; IEEE, 2016.
(18) Nogin, Y.; Detinis Zur, T.; Margalit, S.; Barzilai, I.; Alalouf, O.; Ebenstein, Y.; Shechtman, Y. DeepOM: Single-molecule optical genome mapping via deep learning. *Bioinformatics* **2023**, *39*, btad137.
(19) Goyvaerts, V.; Van Snick, S.; D'Huys, L.; Vitale, R.; Helmer Lauer, M.; Wang, S.; Leen, V.; Dehaen, W.; Hofkens, J. Fluorescent SAM analogues for methyltransferase based DNA labeling. *Chem. Commun.* **2020**, *56*, 3317−3320.
(20) Vranken, C.; Deen, J.; Dirix, L.; Stakenborg, T.; Dehaen, W.; Leen, V.; Hofkens, J.; Neely, R. K. Super-resolution optical DNA Mapping via DNA methyltransferase-directed click chemistry. *Nucleic Acids Res.* **2014**, *42*, No. e50.
(21) Sage, D.; Unser, M. Easy Java programming for teaching image-processing *Proceedings 2001 International Conference on Image Processing (Catal. No.01CH37205)*; IEEE, 2001, pp 298−301.3
(22) Vitale, R.; Cocchi, M.; Biancolillo, A.; Ruckebusch, C.; Marini, F. Class modelling by Soft Independent Modelling of Class Analogy: why, when, how? A tutorial. *Anal. Chim. Acta* **2023**, *1270*, 341304.
(23) Bodor, A.; Bounedjoum, N.; Vincze, G. E.; Erdeiné Kis, Á.; Laczi, K.; Bende, G.; Szilágyi, Á.; Kovács, T.; Perei, K.; Rákhely, G. Challenges of unculturable bacteria: environmental perspectives. *Rev. Environ. Sci. Biotechnol.* **2020**, *19*, 1−22.
(24) Lu, L.; Shin, Y.; Su, Y.; Em Karniadakis, G. Dying ReLU and Initialization: Theory and Numerical Examples. *arXiv* **2019**, arXiv:1903.06733v3.
(25) Darling, A. C. E.; Mau, B.; Blattner, F. R.; Perna, N. T. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Res.* **2004**, *14* (7), 1394−1403.
(26) Ge, Z. Y.; Demyanov, S.; Chen, Z.; Rahil, G. Generative OpenMax for Multi-Class Open Set Classification. *arXiv* **2017**, arXiv:1707.07418.

https://doi.org/10.1021/acsomega.4c09485
ACS Omega 2025, 10, 9224−9232