ACS SENSORS

pubs.acs.org/acssensors Article

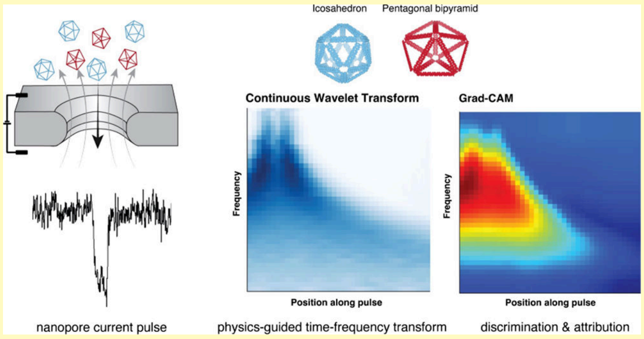
Interpretable Deep Learning for Single-Molecule Nanopore Fingerprinting Using Physics-Guided Preprocessing

Arjav Shah, Xin Kai Lee, Kun Li, Grant A. Knappe, Mark Bathe, George Barbastathis,* and Patrick S. Doyle*

Cite This: https://doi.org/10.1021/acssensors.5c04784 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Rapid and robust molecular fingerprinting is critical in biomanufacturing, diagnostics, and environmental monitoring. Nanopore sensing provides single-molecule readouts as transient ionic current pulses; however, conventional analyses depend on handcrafted features that miss informative structural information. We present an interpretable machine learning framework that operates directly on raw pulses, pairing a physics-guided time−frequency transform with a compact neural classifier and feature-attribution maps. We also include conventional feature-based SVMs and a 1D classifier trained on raw pulses as baselines. On two self-assembled DNA nanostructures of similar size but distinct geometry, for which standard pulse features overlap, the method achieves high accuracy and yields physically consistent attributions that highlight discriminative signal motifs. A matched control without the time−frequency transform clarifies when learned filters suffice versus when physics-guided preprocessing improves reliability, leading to a practical "custom-filter" design principle. The workflow is modular, lightweight, and applicable to pulse-based sensing platforms, including virus and exosome analysis, electrochemical monitoring, and industrial fault detection. By combining accuracy with transparency, it lays the groundwork for deployable sensing platforms in regulated, mission-critical settings.



KEYWORDS: *nanopore sensing, DNA nanostructures, single-molecule fingerprinting, machine learning, explainable AI, wavelet transform, signal processing, convolutional neural networks*

Biomolecular fingerprinting, the identification and characterization of biomolecules based on their unique structural or chemical signatures, is a crucial step across diverse applications, including medical diagnostics, forensic science, food safety, environmental monitoring, epidemiological studies, biotechnology, and quality control for biomanufacturing.[1−4] Given the diversity of biological structures at the nanometer scale, a critical component of these approaches is the discrimination between species of similar size, for example, exosomes, viruses, and lipoproteins.[5−7] In addition, structural similarity, surface markers, and functional mimicry make discrimination between biomolecules challenging. Overcoming these challenges requires a combination of advanced analytical techniques and a multiparameter approach, incorporating various sizes, morphologies, and functional assays, among others, to achieve accurate discrimination.

For example, advanced medicinal therapies such as cell and gene therapies are prone to viral contamination.[8] Sterility testing is critical for detecting harmful contaminants before therapies are released and administered to patients. The gold standard in vitro virus (IVV) assays prescribed for sterility testing by the FDA are comprehensive, cell-based methods that assess key aspects of viral behavior, including replication and infectivity.[9,10] These guidelines require an incubation period of 14−28 days before the cell therapy product can be released for patient administration. This creates a significant bottleneck, delaying treatment for critically ill patients and increasing costs due to product quarantine. The resource-intensive nature of these broad-spectrum IVV assays also limits access in low-resource settings. These barriers highlight the need for rapid, high-throughput, label-free sterility testing methods.

One such proposed technique with desirable characteristics is nanopore sensing. Nanopores have been widely developed and used for nucleic acid sequencing applications.[1,11] More recently, these have been explored for proteomics, diagnostics, and viral detection.[12−15] However, there is little development to increase the speed of these assays further while improving the sensitivity and specificity of detection. *Sensitivity* and *specificity* respectively measure the proportion of actual positives and actual negatives

Received: December 13, 2025
Revised: February 3, 2026
Accepted: February 6, 2026

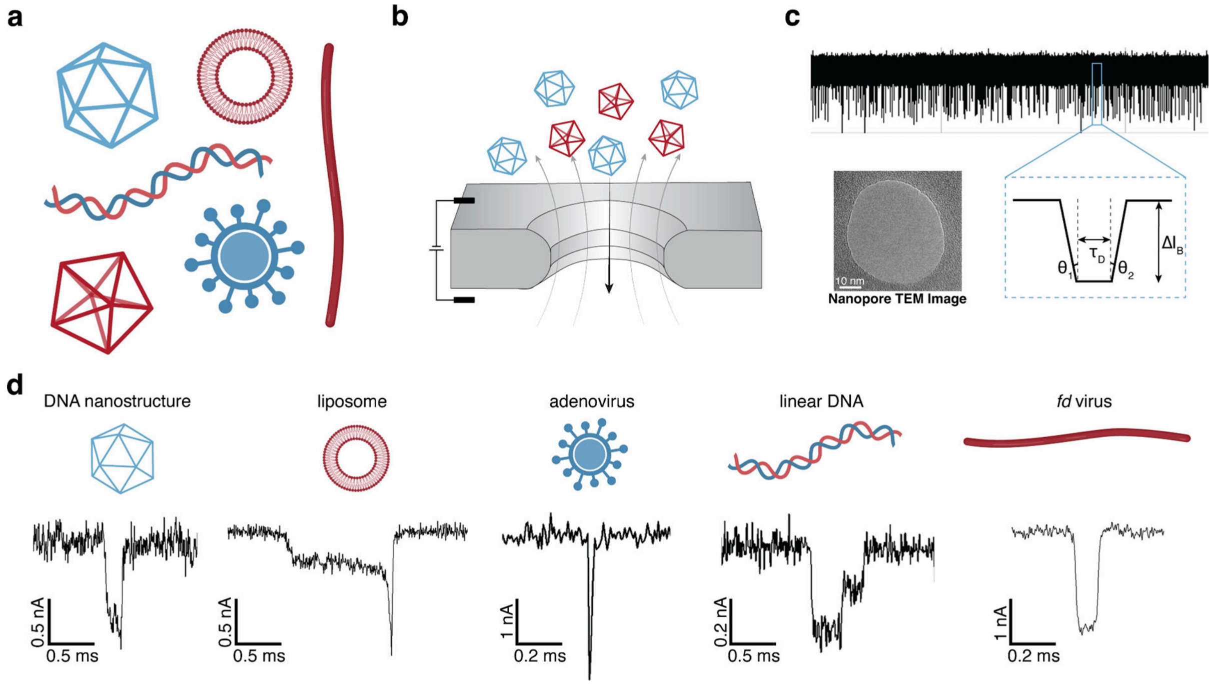


Figure 1. Nanopore sensing of diverse biomolecules and their characteristic current−time (I−t) pulses. (a) Representative nanoscale analytes of similar size: linear DNA, filamentous virus *fd*, liposome/exosome, adenovirus, and DNA nanostructures. (b) Schematic of translocation through a solid-state silicon nitride (SiNx) nanopore under electrophoretic driving forces. (c) Experimental current−time trace from a SiNx pore with a transmission electron microscopy (TEM) image inset; key pulse features used for fingerprinting are indicated—baseline levels $\theta_{1,2}$, dwell time τ_D, and blockade amplitude ΔI_B. (d) Example pulses for each analyte illustrating distinct signatures. Axes: current (nA) versus time (ms).

that are correctly identified by a test. This is critically important, for example, in sterility testing, as the concentration of contaminating viruses is extremely low (~ng/L) and there are several confounding species in the cell culture mixture, including exosomes, viruses, proteins, and cell debris. (Figure 1a). These need to be discriminated from the gene carriers (rAAVs, lentiviruses), lest they lead to a false positive signal. While existing techniques have been explored for discrimination between viruses of varying sizes, it remains challenging to discriminate between species with overlapping size distributions.[16−18]

Resistive current-time pulses from nanopore experiments are exceptionally information-rich, providing detailed insights into the characteristics of individual molecules as they translocate through the nanopore. The nanopore system consists of two reservoirs, containing an electrolytic solution, separated by an insulating, free-standing membrane (Figure 1b). The reservoirs are connected through a pore which is on the order of nanometers in diameter. An electric potential applied across the pore develops a constant ionic current through the pore. Electrophoretic driving forces translocate the charged biomolecules through the pore. The introduction of the biomolecule inside the pore disrupts the flow of ions as it sterically blocks the pore. This generates spike-like drops in an otherwise constant current. These drops are known as ‘resistive pulses’ since their generation is a result of the resistance offered by the particle to ion flow.

These variations in current are a unique *fingerprint* for each molecule and can reveal a wealth of information about its size, shape, charge, and conformation (Figure 1c). The characteristics of these pulses depend on both the biomolecule and pore-related properties, ranging from the diameter and shape of the biomolecule and pore, surface charges, electrolyte concentration, and applied voltage, as seen in Figure 1d.[19−21] The high temporal resolution and sensitivity of nanopore experiments enable an understanding of complex molecular interactions and dynamics, making nanopores powerful tools for applications such as DNA sequencing, molecular fingerprinting, and biomarker detection.[13,22−24]

Each nanopore experiment typically yields thousands of translocation events ($O(10^3)$), generating large volumes of time-resolved ionic current data. While these resistive pulses encode rich biomolecule information, extracting meaningful insights at scale, especially in high-throughput settings involving multiple species, remains challenging. Traditional analyses have often reduced the current−time (I−t) traces to a few low-dimensional features, such as blockade depth (ΔI_B) and translocation duration (τ_D), or their voltage-normalized counterparts (e.g., conductance blockade, ΔG_B). Although these features allow basic classification, they can obscure critical signal characteristics and confound discrimination among similar analytes. Moreover, conventional time-series processing pipelines frequently apply denoising or filtering steps that risk suppressing subtle yet informative pulse features.

To overcome these limitations, machine learning (ML) methods have emerged as powerful tools for real-time, data-driven classification of nanopore signals.[25,26] By leveraging high-dimensional representations, such as time-frequency decompositions and morphological descriptors, modern ML algorithms, including Support Vector Machines (SVM) and deep neural networks, have significantly improved classification performance.[15,16,27−32] These approaches have enabled single-particle identification of closely related viral and bacterial species, including the successful classification of four human coronaviruses using nanopore data.[14] Advanced signal processing techniques, such as wavelet-based denoising and time-frequency transformations, further enhance

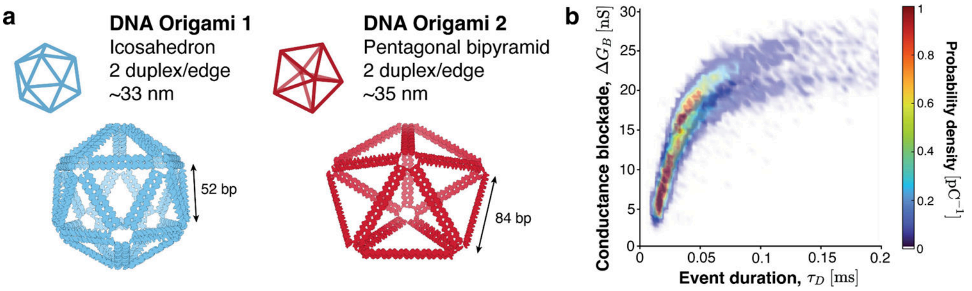


Figure 2. DNA nanostructures pair and baseline separability in standard features. (a) Two self-assembled DNA nanostructures of comparable hydrodynamic size: an icosahedron (~33 nm; 2 duplex/edge; edge length 52 bp) and a pentagonal bipyramid (~35 nm; 2 duplex/edge; edge length 84 bp). Sizes are reported from dynamic light scattering (DLS) measurements. (b) Joint distribution of conductance blockade ΔG_B (nS) versus event duration τ_D (ms) extracted from experimental current−time traces; color denotes probability density. The distributions show substantial overlap, indicating that these standard features alone are insufficient for separability.

model performance by preserving informative features across diverse pulse profiles.[33] Time−frequency analyses have also been applied directly to nanopore events to relate spectral content to single-molecule interactions and dynamics, underscoring the value of spectral views for transient, non-stationary pulses.[34,35] Despite these advances, major challenges remain: model robustness is hindered by poor generalization across experimental conditions; scalability is strained by the growing size and complexity of nanopore datasets; and adaptability is limited by the dependence on narrowly curated training sets.[36,37] Addressing these constraints will be critical to developing universal, high-throughput nanopore platforms capable of sensitive, specific, and broadly deployable molecular fingerprinting.

In this work, we present an integrated framework that combines raw nanopore translocation signals, signal denoising, and machine learning for biomolecular finger-printing. Motivated by prior work, we adopt a time−frequency representation as a physically meaningful front end for classification.[34,35] Our approach leverages Continuous Wavelet Transform (CWT) to decompose ionic current pulses into time-frequency representations, followed by feature extraction using a Convolutional Neural Network (CNN) applied to the resulting two-dimensional scaleograms.[38] This combination enables the model to learn characteristic signal patterns associated with individual biomolecules and classify them based on their 'learned' latent fingerprints. To enhance interpretability, we apply Gradient-weighted Class Activation Mapping (Grad-CAM), which localizes the discriminative regions within the scaleograms that inform the model's predictions.[39] In contrast to prior nanopore−ML pipelines that depend on hand-selected descriptors tied to specific interaction mechanisms (e.g., ΔG_B, τ_D, multilevel states), we pursue a physics-guided time−frequency representation with per-event attribution, avoiding manual feature selection while retaining physical interpretability.[40−42] To our knowledge, this work represents an early step toward integrating explainability into deep learning-based nanopore analytics, offering a path toward more transparent models for molecular classification.

We first describe the dataset and baseline separability ($\Delta G_B - \tau_D$), then present our CWT→CNN pipeline and its performance compared to a feature-based SVM. We then test an ablation that bypasses the CWT and finally, use Grad-CAM to localize discriminative signal regions before discussing robustness and implications.

To validate the performance of our proposed framework, we conducted proof-of-concept experiments using two self-assembled DNA nanostructures of similar size but distinct shape: an icosahedron and a pentagonal bipyramid (Figure 2a). The constructs were computationally designed and experimentally verified to have similar sizes and zeta potential, assayed through DLS (SI Figure S1). They differ in geometry and symmetry, however: the icosahedron is nearly isotropic (icosahedral symmetry), whereas the pentagonal bipyramid is axially anisotropic with fivefold symmetry. DNA nanostructures offer an ideal model system for nanopore sensing due to their well-defined geometry, tunable size, and high monodispersity, allowing for controlled investigations of geometry-dependent signal features.[43−45] Unlike synthetic polymers or heterogeneous biological particles, DNA nanostructures can be precisely designed to isolate specific geometric or structural parameters, making them particularly suited for benchmarking classification algorithms. Experiments were carried out under optimized conditions using solid-state silicon nitride (SiN_x) nanopores (see Experimental Section for details). The resulting datasets exhibit substantial overlap in translocation signals between the two species, making this a stringent benchmark for assessing model discriminability (Figure 2b).

EXPERIMENTAL SECTION

DNA Nanostructure Synthesis

The icosahedral and pentagonal bipyramid DNA nanostructures were fabricated as previously.[46] Briefly, the nanostructures were designed using DAEDALUS with custom sequence ssDNA scaffolds.[44] In a TAE 12 mM $MgCl_2$ buffer, 30 nM scaffold and 150 nM staples were combined and subjected to the following annealing processes: for the icosahedron, 65 °C for 15 min, 61 °C for 90 min, 60 °C for 90 min, 25 °C for 5 min; for the pentagonal bipyramid, 95 °C for 5 min, 80−75 °C at 1 °C per 5 min, 75−30 °C at 1 °C per 15 min, and 30−25 °C at 1 °C per 10 min. After assembly, the nanostructures were purified into PBS using Amicon Ultra centrifugal filters (100 kDa, 2000 × g, 4×) and were stored at 4 °C before use. DNA nanostructure was characterized with agarose gel electrophoresis (1.6 wt % agarose, TAE, 12 mM $MgCl_2$, 60 V, 150 min, room temperature), DLS (at 50 nM), and transmission electron microscopy (TEM)(10 nM DNA nanostructure stained using 2% uranyl acetate on freshly glow-discharged grids). Structural characterization data, including agarose gel electrophoresis, DLS, and TEM, are included in SI Figure S1.

Nanopore Fabrication

Norcada chips (NTDB-B105V122) with a 12 ± 2 nm thickness, square (10 × 10 μm), low-stress SiN_x membrane window centered within a 200 μm-thick silicon frame were used for the pores for the reported experimental data. The pores were fabricated using a Helium Ion

Microscope (HIM), Zeiss Orion Nanofab. The spot size was varied around 3 to achieve an ion current ranging from 2.1 to 2.3 pA while operating at 30 kV of acceleration voltage with a 10 μm aperture. The pore is a standardized circular one with a sub-100 nm diameter. In order to precisely control the size, both X and Y scan pixel spacing were set to 1 nm. A dose of 2 $\times 10^{18}$ ions/cm^2 of He ion beam was applied for etching.

Nanopore Experiments

HIM-fabricated nanopore chips, treated with piranha solution, were mounted between two nanofluidic half-cells filled with the aqueous salt solution. Ag/AgCl electrodes were inserted into each half-cell, and the system was connected to a current amplifier (Axon Axopatch 200B with Digidata 1440B data acquisition system). A constant potential difference of 200 mV was applied across the membrane, and the resulting ionic current was recorded. A grounded Faraday cage used for the apparatus leads to a better ambient electromagnetic screening. Further details of the experimental setup have been previously explained.[20]

Each of the DNA nanostructure samples was diluted to a 2 nM concentration in 1 M KCl, 50 mM Tris, and 10 mM EDTA (pH = 8.3) aqueous solution, and then injected into the reservoir on the cis side.

Event Identification

In order to reduce the continuous noise in the signal and balance noise suppression with signal smoothness, a 5 kHz low-pass FIR filter with a filter order of 40 was applied to each event. The signals obtained from experiments were digitized using a 250 kHz sampling rate. The subsequent analyses of events as described hereafter were carried out using MATLAB scripts. The baseline and amplitude of the signal are identified using a function with a manually set minimum amplitude to ensure that translocation events with prominent peaks are detected. The duration and current values associated with each peak are recorded and, in turn, represent the τ_D and ΔI_B values, respectively. Estimating derivatives using finite differences, the rate of change of the signal is obtained. Based on a defined threshold for the baseline drop (0.2 nA), the approximate start and end points of an event are calculated based on the standard deviation from the threshold. This is a standard protocol used in other published work as well.[20]

It is to be noted that, given the limitations of the instrument, as with any experiment, the signal from some experiments can be distorted due to the filter effect. The cutoff frequency for the preamplifier used in the experiments is 50 kHz (time resolution = 20 μs). Bandwidth limits introduce pulse-shape distortion, but measured-transfer-function analyses and deconvolution indicate that discriminative content is preserved and CWT−CNN remains accurate. Despite the limitations of the instrumentation, the approach to conserve the full pulse trajectory and the associated methodology developed in this work are novel and valuable. For future experiments, a preamplifier with a cutoff frequency of 1−10 MHz (time resolution of 0.1−1 μs) is recommended. With MHz-rate hardware, raw-time models may suffice more often; nonetheless, a time−frequency front end still aids robustness and interpretability for inherently non-stationary, multi-scale nanopore signals.

Algorithms and Dataset

Training, Validation, and Test Dataset. In order to evaluate the model performance in an unbiased manner, we performed a randomized training : validation : test split of the DNA nanostructure pulse signatures with a ratio of 7:2:1. In total, there are 1448 pulse time series for icorsahedron (DNA nanostructure 1) and 1596 pulse time series for pentagonal bipyramid (DNA nanostructure 2). To maintain the ratio of samples between species 1 and 2 in the training, validation, and test datasets, a stratified sampling is performed to generate datasets that are representative of the total population for training, validation, and testing.

Evaluation Metric. We use the F_1-score as the primary evaluation metric to select models that maintain a balanced performance across classes. The F_1-score, defined as the harmonic mean of precision and recall, provides a single measure that balances false positives and false negatives. This choice is motivated by the need for both high sensitivity (recall) and high precision (positive predictive value); accordingly, we report F_1-scores (harmonic mean of precision and recall). This is particularly important in applications such as sterility testing, where misclassification in either direction can have serious consequences. The F_1-score is computed as:

$$F_1 = \frac{2}{\text{precision}^{-1} + \text{recall}^{-1}} \tag{1}$$

$$\text{precision} = \frac{\text{\# of True Positives}}{\text{\# of True Positives} + \text{\# of False Positives}} \tag{2}$$

$$\text{recall} = \frac{\text{\# of True Positives}}{\text{\# of True Positives} + \text{\# of False Nagatives}} \tag{3}$$

where precision measures the ability of a classifier in identifying only relevant data points, while recall measures the ability of a classifier in identifying all relevant data points. The highest possible F_1-score is 100% for a perfect classifier. The F_1-scores for the models are averaged over the 2 classes in the problem.

Training Dynamics. In Figure 6a, we see that the loss function of the training loss is generally decreasing as the training epoch increases. However, the validation loss generally decreases until around epoch 100, plateaus, and increases slightly as training proceeds, which is typical of neural network training. In the beginning, the model is underfitting; thus, the training and validation losses both decrease with training. As training progresses, the training loss continues to decrease, while the validation loss stops decreasing, which indicates that the model has begun to overfit to the training data. We thus select the best neural network weights at the epoch where the F_1-score is maximum on the validation dataset.

Continuous Wavelet Transform (CWT). Continuous Wavelet Transform (CWT) is similar to the Fourier Transform in that the frequency information can be extracted from a time series. However, CWT also retains the time domain information. Similarly, a discrete wavelet transform (DWT) shortens the signal by breaking it down into frequency components, making it computationally more efficient. While DWT is an effective tool for signal denoising, CWT offers advantages over both DWT and FFT, particularly when time-localized frequency information is important.

To obtain a better representation of the time-series data of the pulses created by the DNA nanostructures, we first apply CWT to the isolated signature pulses before performing our classification task. CWT is an analysis technique that extracts local frequency information from a signal, similar to a windowed Fourier transform.[38] CWT can be used to study time series that have nonstationary behavior at many different frequencies that vary over time, while retaining local time information. Using a Fourier transform applied to the entire time-series, local features of the time-series become global in scope in the frequency domain.[47] One common approach to time−frequency analysis of signals is the windowed Fourier transform, which applies a localized windowing function before performing the Fourier transform.[48] While effective, this method uses a fixed window size, which limits its resolution trade-off between time and frequency. The continuous wavelet transform (CWT) addresses this limitation by enabling variable resolution: higher-frequency components are analyzed with finer time resolution, while lower-frequency components benefit from greater frequency resolution.[49] This is achieved using wavelets—basis functions with zero mean and finite support in both time and frequency domains—which allow for adaptive, multi-scale signal decomposition.[50] For contrast, the Fourier transform uses sinusoidal functions as the basis for decomposition. Mathematically, the CWT X_ω of a time-series $x(t)$ is expressed by the integral

$$X_{\omega}(a,b) = \frac{1}{|a|^{\frac{1}{2}}} \int_{-\infty}^{\infty} x(t)\psi^{*}\left(\frac{t-b}{a}\right)\mathrm{d}t, \tag{4}$$

where the scale a is a positive real number, the translational value b is a real number, ψ is the wavelet function, and ψ^* is its complex conjugate. ψ is the source function that is being translated over different times and scaled over different periods to generate different wavelets that resolve the localized large- and small-scale properties of the time-series at the appropriate resolution. In our study, the complex Gaussian wavelet is chosen as the basis function, which is given by

$$\psi(t) = C\frac{\partial}{\partial t}\left(\mathrm{e}^{\mathrm{i}t}\mathrm{e}^{-t^2}\right) \tag{5}$$

where C is a normalization constant. The CWT generates a scaleogram, which rep-resents the intensity of each frequency as a function of time, forming a 2D field. A sample CWT of a DNA nanostructure pulse is shown in Figure 3.

Alternative adaptive time−frequency methods, such as the Hilbert−Huang trans-form (empirical mode decomposition followed by a Hilbert spectrum), are also applicable to non-stationary single-entity signals.[51,52] We did not pursue HHT here because CWT achieved competitive accuracy with fewer tunable parameters, offered stable localization of transient motifs, and proved robust under our noise conditions; by contrast, HHT can be sensitive to mode mixing and endpoint effects and introduces sifting choices that could complicate standardization. A comprehensive comparison of CWT, HHT, and learned front-end filters across larger analyte classes is an important avenue for future work.

Support Vector Machine (SVM). A support vector machine (SVM) is commonly used in classification problems. This linear classifier, which is a part of a broad suite of supervised ML algorithms, finds an optimal line in 2-D or a hyperplane in an N-dimensional space to maximize the separation between different classes[53] (Figure 4a). While multiple hyperplanes may exist, the best decision boundary is found by considering the greatest margin between the data points of both classes, which are on either side of this hyperplane. This enables the algorithm to generalize well to new data and enhance the robustness of classification.

Time-frequency data from the CWT scaleogram is used for classification (Figure 3). While an inherently linear classifier, the use of kernel functions introduces nonlinearity and helps with the classification of nonlinear data, albeit with increased computational complexity. We have explored the performance of 'linear' and 'radial basis function' kernels to capture any non-linearity in the data. The radial basis function, also known as the Gaussian or RBF kernel, has two associated parameters, γ and C. γ determines the influence of samples selected as 'support vectors', with low values implying a far-reaching influence. The C parameter can be thought of as a regularization parameter for SVM. While it is clear from Figure 2b that the data is not linearly separable, at least in the $\Delta G_{\mathrm{B}} - \tau_{\mathrm{D}}$ space, we are working with a wavelet-transformed dataset. As a result, without prior knowledge of the data distribution, an RBF kernel is an intuitive choice given that it can model complex relationships better.

SVMs struggle with datasets where classes significantly overlap and require preprocessed input data to construct a high-dimensional feature space.[36] Generally, SVMs are computationally more expensive to train as compared to other supervised learning classifiers, but less prone to overfitting, while neural networks are generally more scalable and flexible for a wide range of datasets.

Convolutional Neural Network (CNN). After applying CWT on the pulses created by the DNA nanostructures to create a scaleogram, a convolutional neural network (CNN) is used as the classifier to distinguish between the DNA nanostructures. Using this approach, we built a machine-learned neural network model that can discriminate between two DNA nanostructures shown in Figure 2a with similar sizes but different shapes, with overlapping current-time characteristics, with around 93.77% accuracy when applied on the test dataset, which is not used to train and select the hyperparameters of the neural network.

A CNN uses convolutional kernels with trainable weights, which act as custom-made filters to extract meaningful features from a dataset with local correlations, a good example being images. These feature maps are then passed into a dense feed-forward artificial neural network in order to perform the classification task. We first apply a 2D convolutional kernel filter with size 3 × 3, stride 1, and 8 channels to the scaleogram. This is followed by a swish[55] activation function, which acts as the nonlinearity to increase expressivity of the model. A max-pooling operation is per-formed to reduce the spatial dimensions of the feature map while retaining the most important features and reducing their locality dependence. This is followed by another convolution with kernels of size 5 × 5, stride 1, and 16 channels, a swish activation function, and another max-pooling operation. We perform one last convolution operation with kernels of size 5 × 5, stride 1, and 32 channels, a swish operation, and a max-pooling, producing 32 2 × 2 feature maps. The extracted features are then passed into a dense feedforward neural network with 4 hidden layers, with 256 units each. A softmax operation is applied on the neural network output to normalize the output to a probability distribution of the predicted particle classes. A schematic of the CNN is shown in Figure 4b.

The design of the CNN aims to capture the most salient features of the scaleograms while maintaining a high level of time-translation invariance. This is achieved by using many max-pooling operations and minimal padding in order to reduce the size of the feature maps drastically. Our design reduces the scaleogram, which is a 39 × 39-pixel image, into 2 × 2 feature maps, reducing the effect of the time location of the pulse in the original data to a large extent. The severe reduction in dimensions also encourages the model to extract only the most salient features in the scaleogram, likely reducing the effect of overfitting, which is a common issue in deep neural networks with many trainable parameters, where the model fits too closely with the training data and is unable to make accurate predictions outside of the training dataset.

RESULTS AND DISCUSSION

To assess classification under overlapping signal regimes, using two DNA nanostructures of similar size but distinct geometry, we quantify separability with standard pulse features and establish an SVM baseline. We then compare a model trained on a physics-guided time−frequency representation with a 1D neural classifier trained directly on raw pulses, and conclude with accuracy metrics and Grad-CAM visualizations of the discriminative signal regions.

Support Vector Machine

Given the overlap in $\Delta G_{\mathrm{B}} - \tau_{\mathrm{D}}$ (Figure 2b), we first assess a feature-based SVM baseline. Upon training the SVM model with scaleogram input data with two different kernels, we evaluate the performance on the test dataset. As mentioned earlier, data is not linearly separable in the $\Delta G_{\mathrm{B}} - \tau_{\mathrm{D}}$ space. However, given the CWT approach, it is prudent to begin with the simplest approach, i.e., the linear kernel. The model classifies the DNA nanostructures with a 93.44% accuracy (F_1-score = 93.42%) (Figure 5a). Detailed statistics of model results can be found in SI Tables S5−S7. Our results outperform previous studies that have generally reported ≤80% accuracy, 90% at best.[16,29,56]

Several previous approaches have used various transforms on the time-series data for preprocessing. In our case, precise time localization is essential, making Continuous Wavelet Transform (CWT) a suitable choice for feature extraction, given its time-localized frequency content. We also implemented a

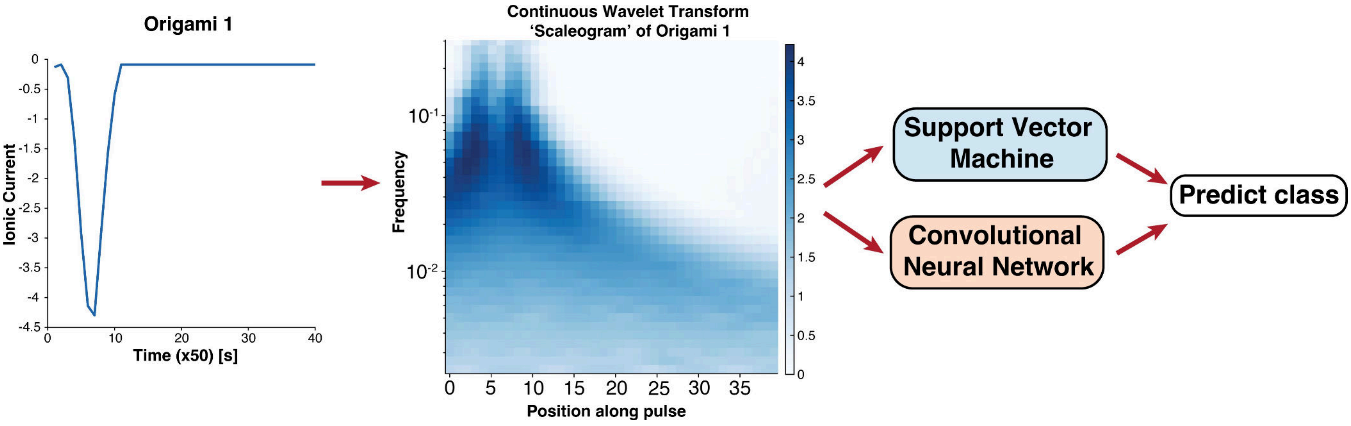


Figure 3. Machine learning pipeline for the classification of the DNA nanostructures using the current−time signal as the particle passes through a nanopore. For a DNA nanostructure pulse signature, a continuous wavelet transform (CWT) is first applied to generate a time−frequency scaleogram. The scaleogram is then passed to two different machine learning frameworks: support vector machine (SVM) and convolutional neural network (CNN), which perform the classification task to infer a species class for the pulse signature. *Position along pulse* refers to the time coordinate along the pulse (×50), expressed in seconds.

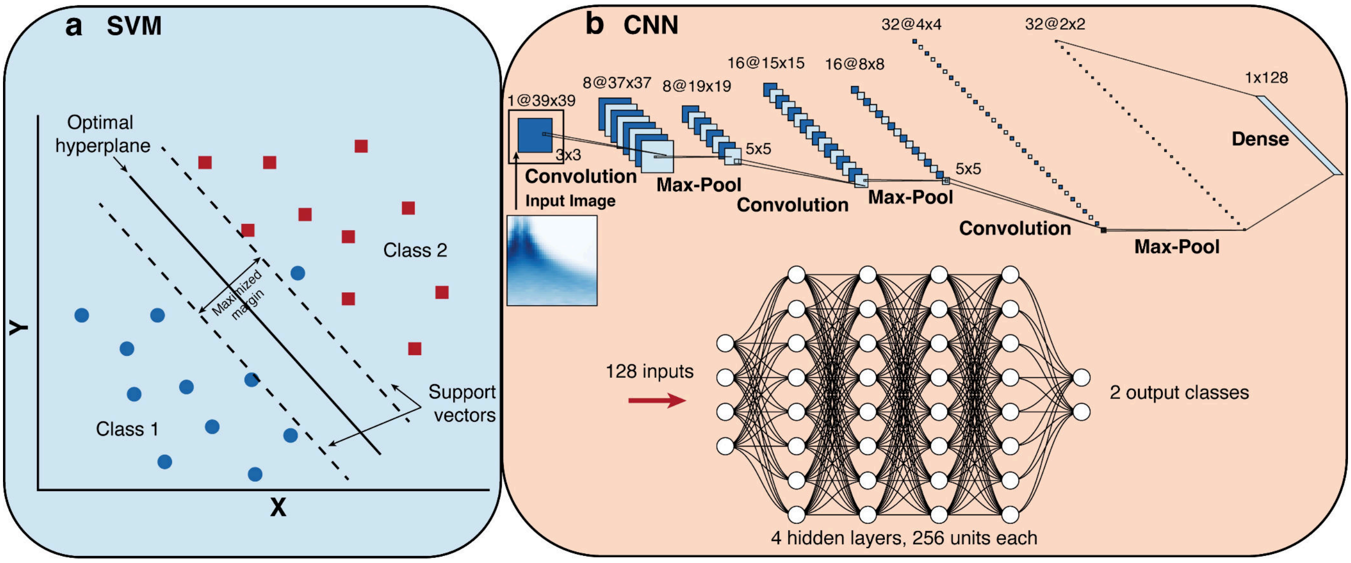


Figure 4. (a) Working principle for SVM. Adapted in part from ref 54. Available under a CC BY-NC-SA license. Copyright 2011 Cambridge University Press. (b) Convolutional neural network (CNN) architecture used for the fingerprinting of DNA nanostructures. Every convolution operation and dense network pass in the CNN is followed by a swish activation function.

support vector machine (SVM) model using raw time-series data without any preprocessing, to benchmark performance without CWT-derived features. The results have been reported in SI Figure S3. The performance drops significantly with an accuracy of 78.69% (F_1-score = 77.97%). Furthermore, we tested the model with an RBF kernel. A hyperparameter search was run to determine the optimal value of γ (SI Table S6). For illustration, we show the confusion matrix for $\gamma = 0.8$ in Figure 5b; the best-performing γ from the sweep is 0.1 (SI Table S6), with test F_1-score = 93.75%. The model performance for optimum γ is comparable to the case with a linear kernel. It must be noted that the training, validation, and testing datasets used for SVM with CWT are identical to the ones created for CNN (discussed in the next section) for fair comparison. While these results are promising and highlight the superior performance of the model with CWT-based preprocessing over previous approaches, SVM can be hard to scale with increasing number of classes (i.e., fingerprinting >2 species), and computationally inefficient to train with large amount of nanopore experimental data and optimize for varying systems (i.e., kernels and associated parameter

Table 1. Validation Performance across Dense-Layer Depth and Width (GELU)[a]

	number of hidden layers			
number of units in each hidden layer	1	2	3	4
64	95.23	94.91	95.56	95.39
128	95.07	95.56	95.40	95.07
256	95.07	95.07	95.07	95.40

[a] Model validation F_1-score (%) for the convolutional neural network (CNN) using a Gaussian Error Linear Unit (GELU) activation function with different hyperparameters. Columns indicate the number of hidden (dense) layers in the classifier head; rows indicate the number of units per hidden layer.

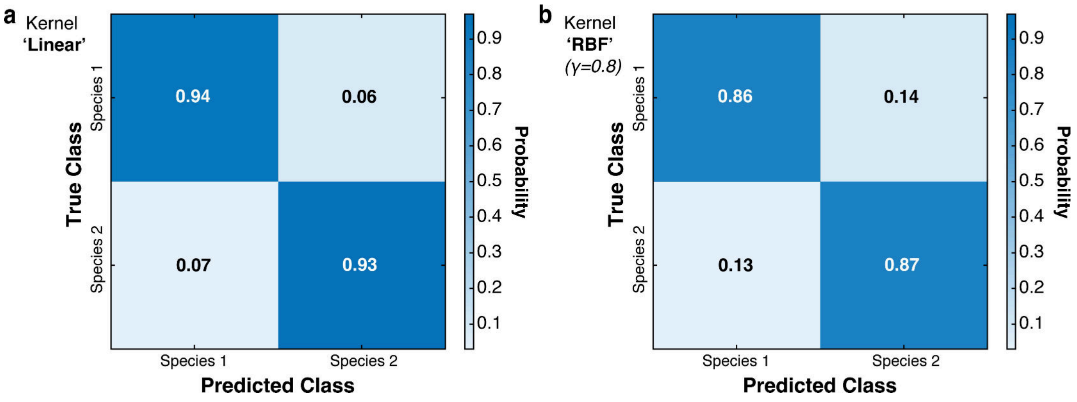


Figure 5. SVM baseline: confusion matrices for linear and RBF kernels. (a) Linear kernel. (b) RBF kernel ($\gamma = 0.8$). Cells show the fraction of test events (row-normalized); color denotes probability. Rows correspond to true classes and columns to predicted classes. The overall test F_1-score is reported in the text.

Table 2. Validation Performance across Dense-Layer Depth and Width (ReLU)[a]

number of units in each hidden layer	number of hidden layers			
	1	2	3	4
64	95.40	95.07	95.23	95.07
128	95.23	95.07	95.07	95.24
256	95.07	94.91	95.23	94.91

[a] Model validation F_1-score (%) for the CNN using a Rectified Linear Unit (ReLU) activation function with different hyperparameters. Columns indicate the number of hidden (dense) layers; rows indicate the number of units per hidden layer.

Table 3. Validation Performance across Dense-Layer Depth and Width (LeakyReLU)[a]

number of units in each hidden layer	number of hidden layers			
	1	2	3	4
64	95.40	94.75	95.07	95.23
128	95.07	95.07	94.91	94.91
256	95.07	94.75	94.91	94.91

[a] Model validation F_1-score (%) for the CNN using a Leaky Rectified Linear Unit (LeakyReLU) activation function with different hyperparameters. Columns indicate the number of hidden (dense) layers; rows indicate the number of units per hidden layer.

Table 4. Validation Performance across Dense-Layer Depth and Width (Swish)[a]

number of units in each hidden layer	number of hidden layers			
	1	2	3	4
64	95.07	95.23	95.56	95.56
128	94.90	95.07	95.40	95.07
256	94.74	94.90	95.39	**95.72**

[a] Model validation F_1-score (%) for the CNN using the Swish activation function with different hyperparameters. Columns give the number of hidden (dense) layers; rows give units per hidden layer. The best configuration is shown in **bold**: 4 hidden layers with 256 units each (95.72%).

choices).[33,57] Because the feature-based SVM struggles in the overlapping regime, we next evaluate a CWT→CNN pipeline.

Convolutional Neural Network

In order to select the best CNN architecture for this classification task, a few hyperparameters were varied, namely the type of activation functions, the number of hidden layers in the dense neural network, and the number of units in each hidden layer. We report F_1-score to balance precision and recall across classes (see Experimental Section). We chose 4 popular activation functions commonly used in machine learning tasks, and varied the number of hidden layers between 1, 2, 3, and 4, and the number of units in each hidden layer between 64, 128, and 256. The models are trained using the Adam optimizer,[58] with a learning rate of $\eta = 3 \times 10^{-4}$ and a cross-entropy loss function. The dataset is batched in chunks of 64, and is trained for 200 epochs each using the Pytorch[59] software package. The best model is selected based on the highest F_1-score when the model is evaluated on the validation dataset. The code developed for the data processing and model training can be found in this GitHub repository. Tables of F_1-scores are shown, with models grouped by the activation function used. Table 1 shows the validation performance with a Gaussian Error Linear Unit (GELU),[60] Table 2 shows the validation performance with a Rectified Linear Unit (ReLU),[61] Table 3 shows the validation performance with a Leaky Rectified Linear Unit (LeakyReLU),[62] while Table 4 shows the validation performance with a Swish[55] activation function.

From Tables tbl1−tbl4, we see that the CNNs in general produce a high level of validation performance with F_1-scores larger than 94.5% for all hyperparameters. We select the hyperparameters of 4 hidden layers with 256 units each and a swish activation function as the final model for the CNN, as it achieves the best validation performance with an F_1-score of 95.72%. Overall, different hyperparameter choices result in comparable accuracy, indicating that the approach is robust and not highly sensitive to specific parameter settings. Evaluating this model on the test dataset yields an F_1-score of 93.77%. Figure 6a shows the loss function of this neural network on the training, validation, and test datasets as a function of epoch, while Figure 6 shows the confusion matrix of the final model on the test dataset.

From Figure 6a, we see that the training/validation losses follow the expected underfit→plateau pattern. We pick the

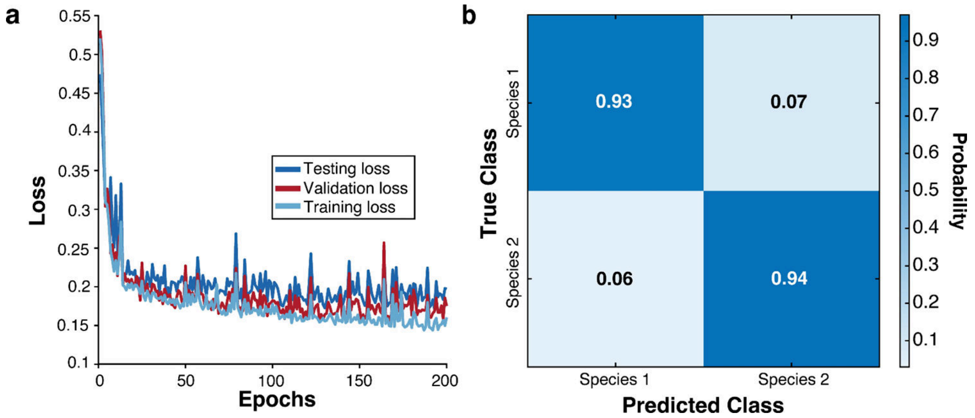


Figure 6. Neural classifier training dynamics and test-set performance. (a) Training, validation, and test losses versus epoch for the selected architecture; the evaluation checkpoint is taken at epoch 43 (peak validation F_1-score). (b) Row-normalized confusion matrix on the test set; rows are true classes and columns are predicted classes, with color indicating probability.

epoch where the F_1-score is maximum on the validation dataset (details in Experimental Section). This is likely the configuration where the model is at its optimum, with good generalization capabilities and balanced precision and recall for both classes. When evaluated on the test dataset in Figure 6b, the model is balanced in its evaluation capabilities of both species and has high accuracy.

Skipping the Continuous Wavelet Transform

To test whether time−frequency filtering is necessary, we trained a 1D-CNN directly on raw pulses. To investigate the effects of bypassing the CWT on the model performance, we also designed a 1D CNN that receives the time series of pulse signatures directly without performing a CWT. In the design of this 1D CNN, we applied the same principles as the 2D CNN, reducing the effect of time translation and promoting latent feature extraction by performing max-pooling aggressively to compress the dimensions of the feature maps. The same training, validation, and test dataset is used in all ML models trained. We find that the model performs slightly worse, with the best validation F_1-score being 95.56% after performing a similar hyperparameter sweep as above. However, the model still performs considerably well. While several previous studies, including Carral et al.,[63] report comparable performance with CNN, those are limited to DNA nucleotide discrimination using a select number of pulse characteristics; more complex datasets from different biomolecules have been harder to parse apart with high accuracy and generalizability.

We speculate that the convolutional architecture that acts on the time series directly is able to create custom filters tailored specifically for this classification task through training and careful consideration of evaluation and selection metrics. These custom filters are able to achieve similar performance to the CWT, which is a universal filter known to be useful in time-series analysis. However, performing a CWT is likely able to reveal local and nonlocal properties of the model slightly more effectively than the 1D CNN, as the CWT creates a 2D image, which is an overrepresentation of the pulse signal, leading to a slight edge. Details about the 1D CNN architecture and training can be found in the Supplementary Information.

Explaining the CNN Decision Using Grad-CAM

Finally, to gain some understanding of the reasoning behind the CNN decision on a particular species, we employ Gradient-weighted Class Activation Mapping (Grad-CAM), a technique in explainable artificial intelligence (XAI).[39] Grad-CAM provides a visual explanation of the region of importance in the input image, which leads the CNN to classify the image as a certain species. This is achieved by using the gradient information flowing into the convolutional layers of interest in the CNN to assign importance values to each neuron for a particular decision. We illustrate the Grad-CAM localization maps of the final CNN on two sample CWT scaleograms of the two different classes, which are the inputs to the CNN, as shown in Figure 7.

Figure 7a,c shows two sample scaleograms of two different classes of DNA nanostructures in the validation dataset. We see that these scaleograms are distinct in their widths along the pulse between the two high-frequency "peaks" in the scaleogram.

This could be a potential discriminator between the two classes. Figure 7b,d shows the Grad-CAM localization maps of the CNN in determining that 7a is of Class 1 while 7c is of Class 2. In the localization maps, red indicates regions of high importance, while blue indicates regions of low importance. From Figure 7b,d, we see that the CNN makes the classification decision primarily based on the gap between the positions of the scaleogram frequency peaks along the pulse. This is perhaps expected and is also a sign that the CNN is not overfitting by picking up spurious correlations in the data; thus, it is likely that this model has good generalization capabilities to unseen data.

■ CONCLUSIONS

Through the use of two DNA nanostructures, this study serves as a proof-of-concept for a novel methodology to distinguish between particles of similar size but different shape. Our approach can be extended to various species, such as liposomes, exosomes, viruses, etc. for discrimination. More broadly, the methodology is application agnostic and can be used widely. While our experiments focus on a binary pair, practical sterility testing typically involves complex mixtures of viruses, extracellular vesicles, and cellular debris. Our event-level model extends directly to K classes via a K-way softmax. Inference is dominated by the CWT front end (independent of K); the final linear layer scales $O(K)$ and adds negligible latency, whereas training scales with data and class count. As class similarity increases, per-class performance may degrade; this can be mitigated by physics-guided preprocessing to emphasize discriminative

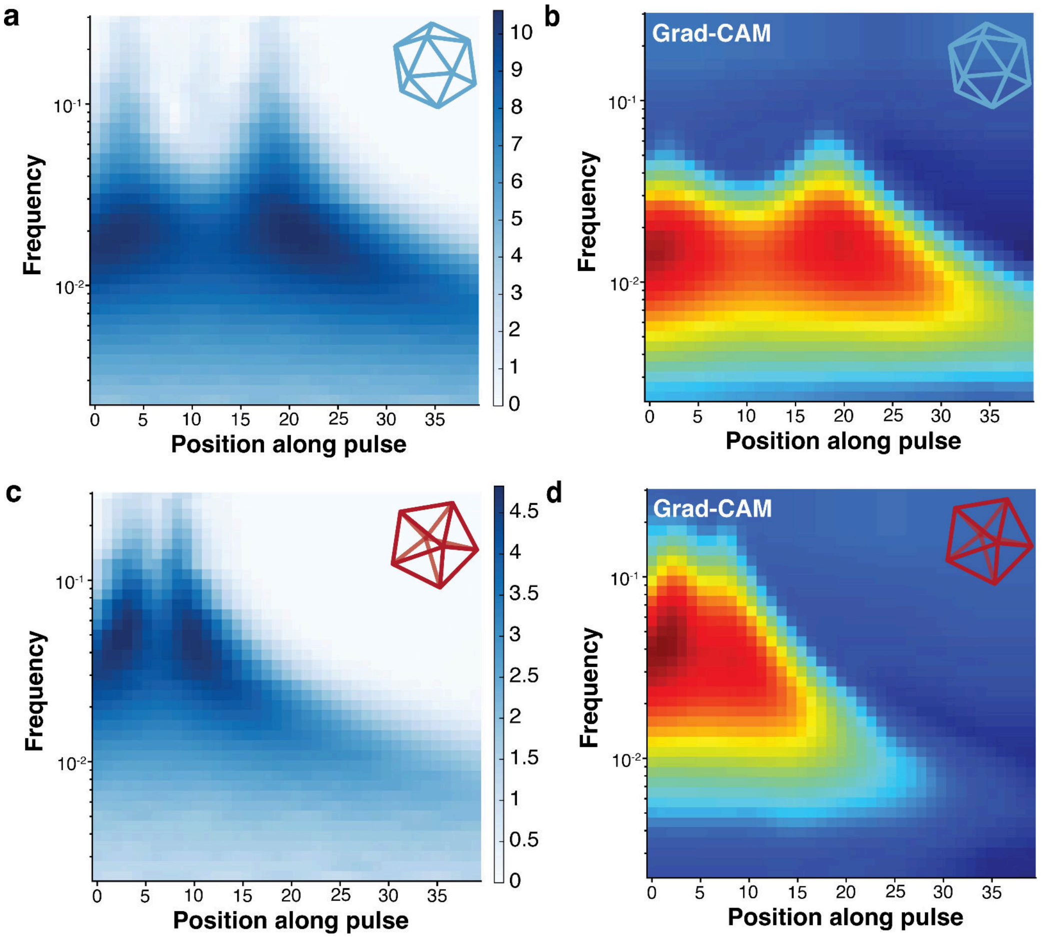


Figure 7. Time−frequency fingerprints and attribution maps for two DNA nanostructures. (a,c) Scaleograms (continuous wavelet transform) of representative validation pulses for the icosahedron (a) and pentagonal bipyramid (c). Axes show time (along the pulse) and frequency (log scale); color denotes normalized magnitude. (b,d) Grad-CAM attribution maps over the corresponding scaleograms, highlighting the regions that most influence the classifier's decision (warm colors = higher attribution, cool colors = lower). Panels a and c show single representative events for illustration; classification is performed over the full dataset, where many pulses exhibit overlapping patterns not separable by eye; hence, the CNN provides automated, scalable decisions, and Grad-CAM localizes the time−frequency evidence for each decision. *Position along pulse* refers to the time coordinate along the pulse (×50), expressed in seconds.

motifs, class-balanced/focal losses with targeted augmentation, and hierarchical labeling with an abstain/OOD option. For complex mixtures, we can classify events individually and infer sample composition from event counts with confidence estimates.

Insights into the decision-making process of the otherwise black-box neural network model are invaluable. It is clear from Figure 6 that our CNN approach is not overfitting the data, ensuring robustness and generalization for a wide range of unseen data. We indeed see that the model is robust across a range of hyperparameters, as reported in Tables tbl1−tbl4, as seen by consistently high F_1-scores on the validation dataset for all hyperparameters. While we interpret the discriminative signatures as primarily geometry-driven, differences in mechanical rigidity (e.g., the pentagonal bipyramid may be more flexible along its short axis under compression) could further contribute to the observed differences in signal (i.e., make the signal differences even greater than due to geometry alone). Targeted perturbations that modulate stiffness (e.g., Mg^{2+} concentration or staple design) would help disentangle geometric from mechanical effects. However, changing Mg^{2+} concurrently alters ionic screening, electrophoretic and electroosmotic forces, and event kinetics, complicating isolation of stiffness. A cleaner approach is to engineer rigidity through DNA nanostructure staple/crossover modifications at fixed buffer conditions and remeasure under identical settings; we identify this targeted follow-up as a priority for future work.

Additionally, although this proof-of-concept uses a single SiN_x pore under fixed conditions to isolate analyte effects, the framework is intended to transfer across devices. For cross-device use, simple normalizations such as expressing blockade depth relative to the open-pore current ($\Delta I/I_0$) or resampling events to a common time grid can standardize inputs. Our pores are fabricated by helium-ion microscopy (HIM), providing tight control over diameter and geometry (see Experimental Section) and reducing inter-chip variability. A systematic multi-pore, multi-chip evaluation is beyond the scope of this work and is an important next step.

While neural differential equations and physics-based ML models[64,65] are beyond the scope of this work, these can be extremely valuable in providing further insights into the decision-making process and improving the accuracy of these models, which is important for critical applications in diagnostics and medicinal therapies. For instance, more recent conductance models like that of Shah et al.[21] and Charron et al.[66] can be used to predict the conductance blockades for a given system, which in turn can be held as a regularizer for the models. More broadly, ML can be

https://doi.org/10.1021/acssensors.5c04784

designed to act as an add-on to quasi-analytical methods instead of supplanting them directly, as done in this work.

In conclusion, this study advances machine learning approaches for nanopore analytics by introducing a structured workflow that combines physically guided preprocessing via Continuous Wavelet Transform (CWT), deep feature extraction using Convolutional Neural Networks (CNNs), and model interpretability through Gradient-weighted Class Activation Mapping (Grad-CAM). This integrated framework not only achieves high classification precision and accuracy for biomolecules based on their nanopore signatures but also improves confidence in model decision-making, an aspect often lacking in conventional black-box ML approaches. More broadly, our results highlight the importance of consciously designing data analysis pipelines with a 'custom filter' approach: either by incorporating domain knowledge through physically motivated transformations like CWT, or by allowing neural networks to learn data-driven filters directly. The key insight is that intentional preprocessing, guided by physical intuition, can substantially enhance model performance, particularly in contrast to feature-poor approaches such as raw SVM classification without signal conditioning.

While the performance gains may be modest in absolute terms, the combination of accuracy, interpretability, and physical grounding makes this approach better positioned for real-world adoption, particularly in applications where reliability and explainability are essential. Translation into regulated environments, such as clinical diagnostics or biomanufacturing quality control, will also require standardized validation procedures and alignment with emerging regulatory frameworks for AI-driven decision-making, potentially guided by the FDA or EMA.

Looking ahead, we anticipate that this framework can serve as a modular, adaptable pipeline for other pulse-based sensing platforms, such as fault detection, electrochemical monitoring, mass spectrometry, or NMR, provided that models are validated rigorously and integrated into hardware−software systems compatible with real-time, high-throughput workflows.[67,68] Key next steps would include cross-pore generalization tests, multi-class expansion, and standardized validation aligned with emerging AI regulatory guidance.

■ ASSOCIATED CONTENT

Data Availability Statement

Pulse data and derived scaleograms are available upon request. The source code for model training, testing, and analysis is available on GitHub (https://github.com/xkykai/DNA-Nanopore-Fingerprinting).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssensors.5c04784.

> Additional model results and comparison, experimental methods, and pulse data, methods including model formulation, network weights, and parameters (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

George Barbastathis – *Singapore-MIT Alliance for Research and Technology Centre, Singapore 138602, Singapore; Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* Email: gbarb@mit.edu

Patrick S. Doyle – *Department of Chemical Engineering, Massachusetts Institute of Technology , Cambridge, Massachusetts 02139, United States; Singapore-MIT Alliance for Research and Technology Centre, Singapore 138602, Singapore;* orcid.org/0000-0003-2147-9172; Email: pdoyle@mit.edu

Authors

Arjav Shah – *Department of Chemical Engineering, Massachusetts Institute of Technology , Cambridge, Massachusetts 02139, United States; Singapore-MIT Alliance for Research and Technology Centre, Singapore 138602, Singapore;* orcid.org/0000-0002-1531-4470

Xin Kai Lee – *The Center for Computational Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-1411-9497

Kun Li – *Singapore-MIT Alliance for Research and Technology Centre, Singapore 138602, Singapore; Department of Physics, National University of Singapore, Singapore 119077, Singapore;* orcid.org/0009-0009-6190-5640

Grant A. Knappe – *Department of Chemical Engineering, Massachusetts Institute of Technology , Cambridge, Massachusetts 02139, United States; Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-5041-2383

Mark Bathe – *Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, United States; Harvard Medical School Initiative for RNA Medicine, Harvard Medical School, Boston, Massachusetts 02115, United States;* orcid.org/0000-0002-6199-6855

Complete contact information is available at: https://pubs.acs.org/doi/10.1021/acssensors.5c04784

Author Contributions

A.S., G.B., and P.S.D. conceived the project. G.A.K. and M.B. designed and synthesized the DNA nanostructures. K.L. performed the experiments. A.S., X.K.L., and K.L. analyzed the data. A.S., X.K.L., and G.B. developed the machine learning model. All authors discussed the results and wrote the manuscript. All authors have approved the final version of the manuscript. A.S., X.K.L., and K.L. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Slaven Garaj for fruitful discussions. This work was supported by the National Research Foundation, Prime Minister's Office, Singapore, under its Campus for Research Excellence and Technological Enterprise (CREATE) program, through the Singapore MIT Alliance for Research and Technology (SMART): Critical Analytics for Manufacturing Personalised-Medicine (CAMP) Interdisciplinary Research Group. A.S. acknowledges support from the MathWorks

Engineering Fellowship. G.A.K. and M.B. gratefully acknowledge funding from the NIH (R01 AI162307) and the ARO-ISN UARC (W911NF2320121).

■ REFERENCES

(1) Wang, Y.; Zhao, Y.; Bollas, A.; Wang, Y.; Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **2021**, *39*, 1348–1365.

(2) Ahmad, M.; Ha, J.-H.; Mayse, L. A.; Presti, M. F.; Wolfe, A. J.; Moody, K. J.; Loh, S. N.; Movileanu, L. A generalizable nanopore sensor for highly specific protein detection at single-molecule precision. *Nat. Commun.* **2023**, *14*, 1374.

(3) Sinha, S. S.; Jones, S.; Pramanik, A.; Ray, P. C. Nanoarchitecture based SERS for biomolecular fingerprinting and label-free disease markers diagnosis. *Acc. Chem. Res.* **2016**, *49*, 2725–2735.

(4) Stollmann, A.; Garcia-Guirado, J.; Hong, J.-S.; Rüedi, P.; Im, H.; Lee, H.; Ortega Arroyo, J.; Quidant, R. Molecular fingerprinting of biological nanopar-ticles with a label-free optofluidic platform. *Nat. Commun.* **2024**, *15*, 4109.

(5) Kashkanova, A. D.; Blessing, M.; Reischke, M.; Baur, J.-O.; Baur, A. S.; Sandoghdar, V.; Van Deun, J. Label-free discrimination of extracellu-lar vesicles from large lipoproteins. *J. Extracell. Vesicle.* **2023**, *12*, 12348.

(6) Cantin, R.; Diou, J.; Bélanger, D.; Tremblay, A. M.; Gilbert, C. Discrimi-nation between exosomes and HIV-1: Purification of both vesicles from cell-free supernatants. *J. Immunol. Methods* **2008**, *338*, 21–30.

(7) Simonsen, J. B. What are we looking at? Extracellular vesicles, lipoproteins, or both? *Circ. Res.* **2017**, *121*, 920–922.

(8) Barone, P. W.; Wiebe, M. E.; Leung, J. C.; Hussein, I. T.; Keumurian, F. J.; Bouressa, J.; Brussel, A.; Chen, D.; Chong, M.; Dehghani, H.; et al. Viral contamination in biologic manufacture and implications for emerging thera-pies. *Nat. Biotechnol.* **2020**, *38*, 563–572.

(9) Center for Biologics Evaluation and Research. *Guidance for industry- character-ization and qualification of cell substrates and other biological materials used in the production of viral vaccines for infectious disease indications*, 2010.

(10) Gombold, J.; Karakasidis, S.; Niksa, P.; Podczasy, J.; Neumann, K.; Richard-son, J.; Sane, N.; Johnson-Leva, R.; Randolph, V.; Sadoff, J.; et al. Systematic evaluation of *in vitro* and *in vivo* adventitious virus assays for the detec-tion of viral contamination of cell banks and biological products. *Vaccine* **2014**, *32*, 2916–2926.

(11) *How nanopore sequencing works*. https://nanoporetech.com/platform/technology.

(12) Dorey, A.; Howorka, S. Nanopore DNA sequencing technologies and their applications towards single-molecule proteomics. *Nat. Chem.* **2024**, *16*, 314–334.

(13) Ying, Y.-L.; Hu, Z.-L.; Zhang, S.; Qing, Y.; Fragasso, A.; Maglia, G.; Meller, A.; Bayley, H.; Dekker, C.; Long, Y.-T. Nanopore-based technologies beyond DNA sequencing. *Nat. Nanotechnol.* **2022**, *17*, 1136–1146.

(14) Taniguchi, M.; Minami, S.; Ono, C.; Hamajima, R.; Morimura, A.; Hamaguchi, S.; Akeda, Y.; Kanai, Y.; Kobayashi, T.; Kamitani, W.; et al. Combining machine learning and nanopore construction creates an artificial intelligence nanopore for coronavirus detection. *Nat. Commun.* **2021**, *12*, 3726.

(15) Arima, A.; Harlisa, I. H.; Yoshida, T.; Tsutsui, M.; Tanaka, M.; Yokota, K.; Tono-mura, W.; Yasuda, J.; Taniguchi, M.; Washio, T.; et al. Identifying single viruses using biorecognition solid-state nanopores. *J. Am. Chem. Soc.* **2018**, *140*, 16834–16841.

(16) Arima, A.; Tsutsui, M.; Harlisa, I. H.; Yoshida, T.; Tanaka, M.; Yokota, K.; Tonomura, W.; Taniguchi, M.; Okochi, M.; Washio, T.; et al. Selective detections of single-viruses using solid-state nanopores. *Sci. Rep.* **2018**, *8*, 16305.

(17) Akhtarian, S.; Miri, S.; Doostmohammadi, A.; Brar, S. K.; Rezai, P. Nanopore sensors for viral particle quantification: current progress and future prospects. *Bioengineered* **2021**, *12*, 9189–9215.

(18) Cassedy, A.; Parle-McDermott, A.; O'Kennedy, R. Virus detection: A review of the current and emerging molecular and immunological methods. *Front. Mol. Biosci.* **2021**, *8*, No. 637559.

(19) Shi, W.; Friedman, A. K.; Baker, L. A. Nanopore sensing. *Anal. Chem.* **2017**, *89*, 157–188.

(20) Li, K.; Shah, A.; Sharma, R. K.; Adkins, R.; Marjanovic, T.; Doyle, P. S.; Garaj, S. Metrology of individual small viruses. *Adv. Mater. Interfaces* **2023**, *10*, No. 2300385.

(21) Shah, A.; Pathak, S.; Li, K.; Garaj, S.; Bazant, M. Z.; Gupta, A.; Doyle, P. S. A universal approximation for conductance blockade in thin nanopore membranes. *Nano Lett.* **2024**, *24*, 4766–4773.

(22) He, L.; Tessier, D. R.; Briggs, K.; Tsangaris, M.; Charron, M.; McConnell, E. M.; Lomovtsev, D.; Tabard-Cossa, V. Digital immunoassay for biomarker concentra-tion quantification using solid-state nanopores. *Nat. Commun.* **2021**, *12*, 5348.

(23) Alibakhshi, M. A.; Halman, J. R.; Wilson, J.; Aksimentiev, A.; Afonin, K. A.; Wanunu, M. Picomolar fingerprinting of nucleic acid nanoparticles using solid-state nanopores. *ACS Nano* **2017**, *11*, 9701–9710.

(24) Confederat, S.; Sandei, I.; Mohanan, G.; Wälti, C.; Actis, P. Nanopore fin-gerprinting of supramolecular DNA nanostructures. *Biophys. J.* **2022**, *121*, 4882.

(25) Zhang, Y.; Wright, M. A.; Saar, K. L.; Challa, P.; Morgunov, A. S.; Peter, Q. A. E.; Devenish, S.; Dobson, C. M.; Knowles, T. P. Machine learning-aided protein identification from multidimensional signatures. *Lab Chip* **2021**, *21*, 2922–2931.

(26) Rickert, C. A.; Lieleg, O. Machine learning approaches for biomolec-ular, biophysical, and biomaterials research. *Biophys. Rev.* **2022**, 3, No. 021306.

(27) Hattori, S.; Sekido, R.; Leong, I. W.; Tsutsui, M.; Arima, A.; Tanaka, M.; Yokota, K.; Washio, T.; Kawai, T.; Okochi, M. Machine learning-driven electronic identifications of single pathogenic bacteria. *Sci. Rep.* **2020**, *10*, 15525.

(28) Tsutsui, M.; Takaai, T.; Yokota, K.; Kawai, T.; Washio, T. Deep learning-enhanced nanopore sensing of single-nanoparticle translocation dynamics. *Small Methods* **2021**, *52*, No. 100191.

(29) Arima, A.; Tsutsui, M.; Yoshida, T.; Tatematsu, K.; Yamazaki, T.; Yokota, K.; Kuroda, S.; Washio, T.; Baba, Y.; Kawai, T. Digital pathology platform for respiratory tract infection diagnosis via multiplex single-particle detections. *ACS Sens.* **2020**, *5*, 3398–3403.

(30) Leong, Y. X.; Tan, E. X.; Leong, S. X.; Lin Koh, C. S.; Thanh Nguyen, L. B.; Ting Chen, J. R.; Xia, K.; Ling, X. Y. Where nanosensors meet machine learning: Prospects and challenges in detecting Disease X. *ACS Nano* **2022**, *16*, 13279–13293.

(31) Pinholt, H. D.; Bohr, S.-R.; Iversen, J. F.; Boomsma, W.; Hatzakis, N. S.; Novo, D. Single-particle diffusional fingerprinting: A machine-learning framework for quantitative analysis of heterogeneous diffusion. *Biophys. Comput. Biol.* **2021**, *118*, No. e2104624118.

(32) Tsutsui, M.; Yoshida, T.; Yokota, K.; Yasaki, H.; Yasui, T.; Arima, A.; Tonomura, W.; Nagashima, K.; Yanagida, T.; Kaji, N.; et al. Discriminating single-bacterial shape using low-aspect-ratio pores. *Sci. Rep.* **2017**, *7*, 17371.

(33) Shekar, S.; Chien, C.-C.; Hartel, A.; Ong, P.; Clarke, O. B.; Marks, A.; Drndic, M.; Shepard, K. L. Wavelet denoising of high-bandwidth nanopore and ion-channel signals. *Nano Lett.* **2019**, *19*, 1090–1097.

(34) Liu, S.-C.; Li, M.-X.; Li, M.-Y.; Wang, Y.-Q.; Ying, Y.-L.; Wan, Y.-J.; Long, Y.-T. Measuring a frequency spectrum for single-molecule interactions with a confined nanopore. *Faraday Discuss.* **2018**, *210*, 87–99.

(35) Fu, Y.-H.; Li, X.; Ma, L.; Wan, Y.-J.; Zhang, L.-M.; Ying, Y.-L.; Long, Y.-T. Exploring the single-molecule transient interactions with nanopore frequency spectrum. *J. Phys. Chem. C* **2024**, *128*, 1110.

(36) Wen, C.; Dematties, D.; Zhang, S. L. A guide to signal processing algorithms for nanopore sensors. *ACS Sens.* **2021**, *6*, 3536–3555.

(37) Dematties, D.; Wen, C.; Pérez, M. D.; Zhou, D.; Zhang, S. L. Deep learning of nanopore sensing signals using a bi-path network. *ACS Nano* **2021**, *15*, 14419–14429.

(38) Torrence, C.; Compo, G. P. A practical guide to wavelet analy-sis. *Bull. Am. Meteorol. Soc.* **1998**, *79*, 61–78.

(39) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based local-ization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359.

(40) Li, M.-Y.; Ying, Y.-L.; Yu, J.; Liu, S.-C.; Wang, Y.-Q.; Li, S.; Long, Y.-T. Revisiting the origin of nanopore current blockage for volume difference sensing at the atomic level. *JACS Au* **2021**, *1*, 967–976.

(41) Liu, W.; Zhu, Q.; Yang, C.-N.; Fu, Y.-H.; Zhang, J.-C.; Li, M.-Y.; Yang, Z.-L.; Xin, K.-L.; Ma, J.; Winterhalter, M.; et al. Single-molecule sensing inside stereo- and regio-defined hetero-nanopores. *Nat. Nanotechnol.* **2024**, *19*, 1693–1701.

(42) Gao, F.; Wang, J.-H.; Ma, H.; Xia, B.; Wen, L.; Long, Y.-T.; Ying, Y.-L. Identification of oligosaccharide isomers using electrostatically asymmetric OmpF nanopore. *Angew. Chem. Int. Ed.* **2025**, *64*, No. e202422118.

(43) Rothemund, P. W. K. Folding DNA to create nanoscale shapes and pat-terns. *Nature* **2006**, *440*, 297–302.

(44) Veneziano, R.; Ratanalert, S.; Zhang, K.; Zhang, F.; Yan, H.; Chiu, W.; Bathe, M. Designer nanoscale DNA assemblies programmed from the top down. *Science* **2016**, *352*, 1534.

(45) Dey, S.; Fan, C.; Gothelf, K. V.; Li, J.; Lin, C.; Liu, L.; Liu, N.; Nijen-huis, M. A. D.; Saccà, B.; Simmel, F. C.; et al. DNA origami. *Nat. Rev. Methods Prime.* **2021**, *1*, 13.

(46) Ofoegbu, P. C.; Knappe, G. A.; Romanov, A.; Draper, B. E.; Bathe, M.; Jarrold, M. F. Charge detection mass spectrometry enables molecular characterization of nucleic acid nanoparticles. *ACS Nano* **2024**, *18*, 23301–23309.

(47) Wirsing, K. Time frequency analysis of wavelet and fourier transform. In *Wavelet Theory*, Mohammady, S., Ed.; IntechOpen, 2021.

(48) Gabor, D. Theory of communication. Part 3: Frequency compression and expan-sion . *J. Inst. Electric. Eng. - Part III: Radio Commun. Eng.* **1946**, *93*, 445–457.

(49) Daubechies, I. The wavelet transform, time-frequency localization and signal analysis. *IEEE Trans. Inf. Theory* **1990**, *36*, 961–1005.

(50) Farge, M. Wavelet transforms and their applications to turbulence. *Annu. Rev. Fluid Mech.* **1992**, *24*, 395–458.

(51) Huang, N. E.; Shen, Z.; Long, S. R.; Wu, M. C.; Shih, H. H.; Zheng, Q.; Yen, N.-C.; Tung, C. C.; Liu, H. H. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. A: Math. Phys. Eng. Sci.* **1998**, *454*, 903–995.

(52) Li, X.; Fu, Y.-H.; Wei, N.; Yu, R.-J.; Bhatti, H.; Zhang, L.; Yan, F.; Xia, F.; Ewing, A. G.; Long, Y.-T.; et al. Emerging data processing methods for single-entity elec-trochemistry. *Angew. Chem. Int. Ed.* **2024**, *63*, No. e202316551.

(53) Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer Science & Business Media, 2008.

(54) Mourao-Miranda, J.; Reinders, A. a. T. S.; Rocha-Rego, V.; Lappin, J.; Rondina, J.; Morgan, C.; Morgan, K. D.; Fearon, P.; Jones, P. B.; Doody, G. A.; et al. Individualized prediction of illness course at the first psychotic episode: a sup-port vector machine MRI study. *Psychol. Med.* **2012**, *42*, 1037–1047.

(55) Ramachandran, P.; Zoph, B.; Le, Q. V. Searching for activation functions. *arxiv* 2017. 10.48550/arXiv.1710.05941.

(56) Arima, A.; Tsutsui, M.; Washio, T.; Baba, Y.; Kawai, T. Solid-state nanopore platform integrated with machine learning for digital diagnosis of virus infec-tion. *Anal. Chem.* **2021**, *93*, 215–227.

(57) Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* **2020**, *408*, 189–215.

(58) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980* 2017. 10.48550/arXiv.1412.6980.

(59) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. PyTorch: An imperative style, high-performance deep learning library. *arXiv:1912.01703* 2019. 10.48550/arXiv.1912.01703.

(60) Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* 2016. 10.48550/arXiv.1606.08415.

(61) Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In *JMLR Workshop and Conference Proceedings*, 2011; pp 315−323.

(62) Maas, A.; Hannun, A.; Ng, A. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the 30th International Conference on Ma- chine Learning, Atlanta, Georgia, USA*, 2013.

(63) Carral, D.; Ostertag, M.; Fyta, M. Deep learning for nanopore ionic current blockades. *J. Chem. Phys.* **2021**, *154*, No. 044111.

(64) Raissi, M.; Perdikaris, P.; Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involv-ing nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707.

(65) Rackauckas, C.; Ma, Y.; Martensen, J.; Warner, C.; Zubov, K.; Supekar, R.; Skinner, D. ; Universal differential equations for scientific machine learning. *axXiv* 2020. 10.48550/arXiv.2001.04385.

(66) Charron, M.; Roelen, Z.; Wadhwa, D.; Tabard-Cossa, V. Improved conductance blockage modeling of cylindrical nanopores, from 2D to thick membranes. *Nano Lett.* **2024**, *24*, 10527–10533.

(67) Chiron, L.; van Agthoven, M. A.; Kieffer, B.; Rolando, C.; Delsuc, M.-A. Effi-cient denoising algorithms for large experimental datasets and their applications in Fourier transform ion cyclotron resonance mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 1385–1390.

(68) Samantaray, Y.; Cogswell, D. A.; Cohen, A. E.; Bazant, M. Z. Electrochemically resolved acoustic emissions from Li-ion batteries. *ChemRxiv* 2025. 10.26434/chemrxiv-2025-r7vwq.