A ML Framework for Genetic Sequence Identification Using 2D Electrical Conductance Probability Distributions from Mixed Data Sets

Yiren Wang, Hongning Wang, Arindam K. Das, and M. P. Anantram, *Senior Member, IEEE*

***Abstract*—Genetic sequence identification from electrical characterization of single molecules has emerged as a promising alternative to traditional approaches. Since electrical data on single molecules is extremely noisy due to the limitations of even state-of-the-art approaches, achieving high detection rates is challenging, particularly when the task involves being able to distinguish a sequence from its single base-pair mismatches. To address this issue, we propose an architecture based on combining a convolutional neural network with an ensemble learning method, XGBoost. In addition, four different input feature representations are considered, 1D conductance probability distributions and 2D conductance versus distance probability distributions which can be viewed as images, with or without averaging over the experimental parameters. The *with averaging* case corresponds to feature matrices derived from mixed datasets. We find that 2D probability distributions are helpful with respect to classifier accuracy, but averaged conductance probability distributions are much more impactful and significantly enhance prediction accuracy. Our quantitative analysis of multiple sequences shows an impressive performance increase of approximately 10% for all sequences. While the basis of our analysis is conductance data of DNA strands and their single base-pair mismatches, our method is generally applicable to other single-molecule identification based on their conductance.**

***Index Terms*—Convolutional neural networks, decision trees, DNA, ensemble learning, genetic mutations, probability distribution.**

I. Introduction

A DNA sequence, a complex strand of nucleotides, is the blueprint of every living organism, including viruses. In the past several decades, DNA sequencing techniques have undergone significant development, leading to profound advancements in our understanding of the DNA molecule and its applications in various disciplines, such as molecular biology, genomics, evolutionary biology, virology, medicine, forensics, medicine, and forensics [1], [2]. Within these disciplines, a key challenge is the accurate identification and classification of genetic sequences and their mutations, particularly those involving single-base pair mismatches [3]. The ability to accurately identify mutations with single/few base-pair mismatches holds particular significance in the context of viral testing, as evidenced during the recent COVID-19 pandemic.

In recent years, several researchers have explored/proposed single molecule all-electronic methods for DNA/RNA sequencing [4], [5], [6], [7], [8], [9] and chemical [10], [11], [12], [13], [14], [15], [16], [17] identification. Broadly speaking, these methods probe the electrical properties of a single molecule by subjecting it to an external voltage source and recording the current (or equivalently, the conductance of the molecule) as the experiment is conducted. Within this paradigm, the identification of sequences rests on the hypothesis that different sequences exhibit different conductance properties, specifically, conductance probability distributions. Such distributions can, therefore, be viewed as "electronic fingerprints" of molecules. Statistical or machine learning methods can then be employed to design classifiers which can produce highly accurate results in near real-time [18], [19]. However, accurate identification of genetic sequences based on their electronic fingerprints can be extremely challenging since even the state-of-the-art methods are inherently stochastic, necessitating multiple experimental passes to collect a "corpus" of data for every molecule [20], [21], [22], [23], [24], [25], [26]. Of course, conductance probability distributions derived from a corpus of data benefit from noise averaging, resulting in more accurate descriptors of the underlying true electronic fingerprints. Nevertheless, from a usability perspective, the holy grail remains highly accurate and robust identification based on data collected from a single experimental pass.

In this paper, we demonstrate that accurate classifiers can be built utilizing the electronic fingerprints of multiple COVID-19 variants and their single base-pair mismatches. The data used in this paper (published in reference [27]) was collected using the single molecule break junction approach under a range of experimental parameters: applied voltage bias, current

Received 4 June 2024; revised 14 January 2025; accepted 24 January 2025. Date of publication 29 January 2025; date of current version 3 April 2025. This work was supported by the National Science Foundation under Grant GCR 2317843, Grant FMRG 2328217, and Grant FuSe-TG 2235294. *(Yiren Wang and Hongning Wang contributed equally to this work.) (Corresponding authors: M. P. Anantram; Arindam K. Das)*

Yiren Wang and Hongning Wang are with the Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98115 USA (e-mail: ethanwyr@uw.edu; howang1196@gmail.com).

Arindam K. Das is with the Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98115 USA, and also with the Department of Computer Science and Electrical Engineering, Eastern Washington University, Cheney, WA 99004 USA (e-mail: arindam@uw.edu).

M. P. Anantram is with the Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98115 USA (e-mail: anantmp@uw.edu).

The code implementation is available at https://github.com/ethanwyr/SMBJClassifier

This article has supplementary downloadable material available at https://doi.org/10.1109/TCBBIO.2025.3536282, provided by the authors.

Digital Object Identifier 10.1109/TCBBIO.2025.3536282

2998-4165 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University of Washington Libraries. Downloaded on May 14,2025 at 23:20:13 UTC from IEEE Xplore. Restrictions apply.

amplifier sensitivity, distance between electrodes, and ramp rate (explained in the Methods section). As a benchmarking step, we first applied the XGBoost classifier to one-dimensional (1D) distributions *without averaging* [18], which yielded low accuracies (83.96%). We then develop a new methodology involving two-dimensional (2D) conductance distributions and an architecture that integrates a convolutional neural network and an XGBoost classifier. Additionally, we investigated *averaged* conductance distributions, where the averaging is over the experimental parameters. Our primary research objective was to explore whether these variations could utilize electronic fingerprints to detect sequences with a higher accuracy. Our secondary objectives were to obtain insights into the number of experimental passes needed to construct reliable conductance distributions and perform a statistical analysis to determine the role of the applied voltage in determining the accuracy of sequence identification. In the Results section, the developed method is used to show that: (i) averaged conductance distributions provide an across-the-board enhancement in classifier accuracy (92.09%), (ii) adoption of 2D distributions can be beneficial for some sequences (93.27%), (iii) twenty experimental passes are adequate to yield classifiers which exceed 90% accuracy for six of the ten sequences considered, but at least fifty passes are needed for four other sequences, and (iv) data collected from experiments conducted at lower voltage biases tend to produce better classifiers.

II. Method

A. Datasets

The dataset used contains several thousands of current traces corresponding to ten unique 12 base-pair (bp) short DNA segments from one of three variants of the SARS-CoV-2 genome [27], as detailed in Table S.I of Supplemental Material. These variants are the Alpha variant B.1.1.7, Beta variant B.1.351, and the Delta variant B.1.167 [28]. Of the ten DNA sequences, three are perfectly matched with the target Alpha, Beta, and Delta variants, which we refer to as Alpha_PM, Beta_PM, and Delta_PM (perfect match) sequences, respectively. The others correspond to sequences of the same length but with a single base pair mismatch, which we refer to as mismatches (MM). The mismatch sequences, denoted by Alpha_MM1, Beta_MM1, and Delta_MM1, are derived from the wild-type SARS-CoV-2 virus, which also exists in some of the unmutated segments of Alpha, Beta, and Delta variants. The mismatch sequences Alpha_MM2, Beta_MM2, and Delta_MM2 are hypothetical sequences that could produce the same virus protein as the PM sequences [27]. Table S.I of Supplemental Material shows the exact mutation locations for these sequences. Detailed rationale and conductance data can be found in [27].

The current traces are sensitive to the choice of experimental parameters such as the applied voltage bias, ramp rate (i.e., the rate at which the two electrical contacts are separated), and the sensitivity of the current amplifier used to measure conductance. The voltage bias was chosen from the set $\{0.03,\ 0.05,\ 0.10,\ 0.15,\ 0.20,\ 0.30\}$ V, the ramp rate applied was from the set $\{3,\ 5,\ 10,\ 20\}$ V/s, and the current amplifier sensitivity is either 10 nA/V or 1 nA/V (better resolution). As shown in Table S.I of Supplemental Material, this results in a total of 39 experimental datasets of conductance traces for the ten COVID-19 sequences. We will discuss the impact of voltage bias on classifier accuracy in detail later.

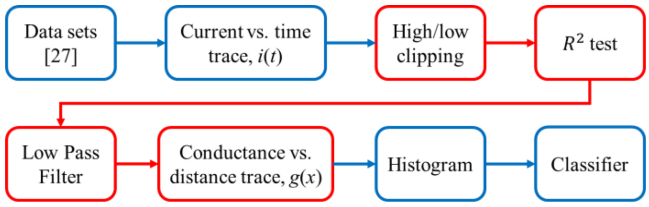


Fig. 1. The sequence of pre-processing steps (shown within red boxes) for converting raw current traces to conductance traces.

B. Pre-Processing Procedure

A series of pre-processing steps were carried out to convert the experimentally obtained 'raw' current traces to suitable conductance traces, which are then used to derive the conductance histograms for training our ML models. Fig. 1 summarizes the pre-processing steps. Fig. 2(a) shows a DNA between the two electrodes, where the applied bias drives the current flow, the ramp rate determines how fast the distance between the electrodes increases, and the preamplifier can also be varied in the experiments. The recorded current values are the outputs of a preamplifier, which depends on the selection of a current amplifier of either 1 nA/V or 10 nA/V sensitivity [27]. Since an amplifier of 1 nA/V sensitivity has a noise floor of 1 pA and an upper limit of 10 nA, we first clipped all current traces recorded with such an amplifier to the range [1 pA – 10 nA]. Similarly, since a 10 nA/V amplifier has a noise floor of 10 pA and an upper limit of 100 nA, we clipped all current traces recorded with such an amplifier to the range [10 pA – 100 nA].

Next, we fit an exponential regression model, $y = a \cdot e^{-b \cdot x},\ b > 0$, to each high/low clipped current trace and used the R^2 statistic from the fit to accept/reject the trace. Since molecular binding between the tip and the substrate (see Fig. 2(a)) on every run of the experimental procedure is not guaranteed, it is possible for "spurious" current/conductance traces to be recorded, which need to be filtered out. Since these spurious traces are predominantly exponentially decaying (in contrast, traces with molecular binding exhibit plateaus), we adopted an R^2 test to eliminate such traces (invalid data). Specifically, if the computed R^2 statistic is greater than some chosen threshold, β, we reject the trace as 'invalid' (experiment conducted with no molecular binding); otherwise, the trace is accepted (see Fig. 2(b) for comparison). We chose a threshold of $\beta = 0.95$ to achieve a balance between filtering out invalid data and retention of adequate data for training and validation. Even though the standard error of regression is a better choice for nonlinear regression models, we determined through extensive experimentation that the R^2 statistic works well for our purposes. We started with about 2000–5000 traces for each of the 39 datasets. For the Alpha variant, we found that approximately 60–85% of the data were accepted at $\beta = 0.95$. There were two

Authorized licensed use limited to: University of Washington Libraries. Downloaded on May 14,2025 at 23:20:13 UTC from IEEE Xplore. Restrictions apply.

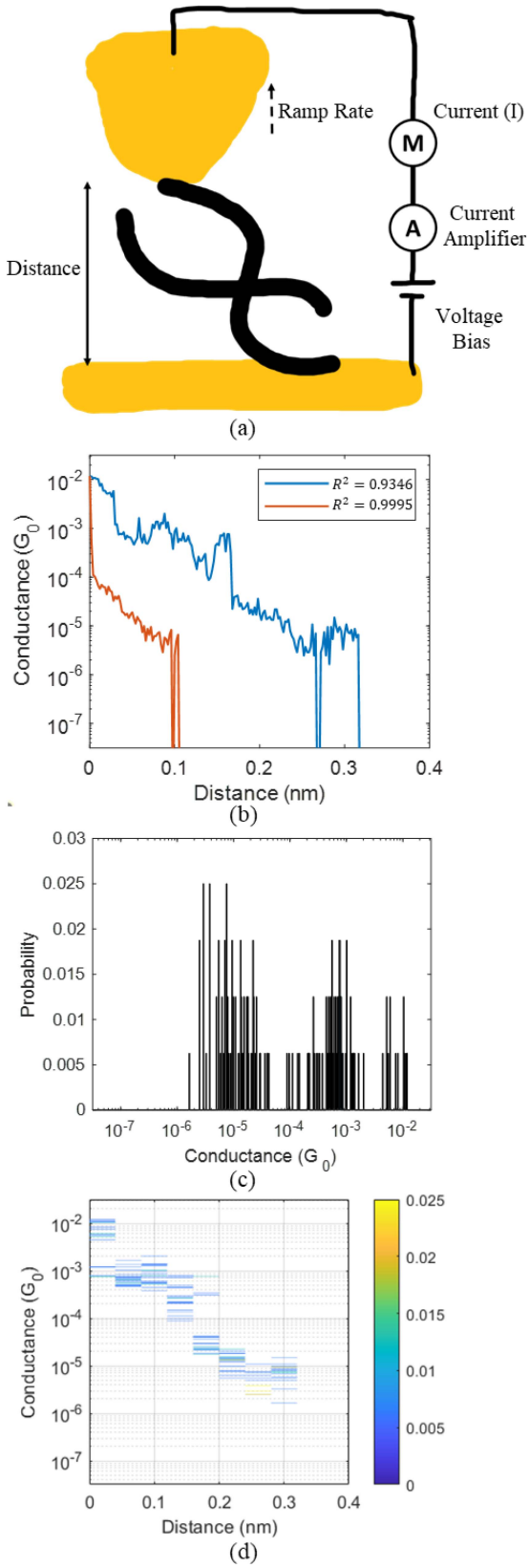


Fig. 2. (a) Schematic of the experimental method. (b) A comparison between two conductance versus distance traces, a valid trace (blue) vs. an invalid trace (red). (c) 1D conductance histogram and (d) 2D conductance histogram of valid trace.

exceptions: E7 corresponding to Alpha_MM2 (42% accepted) and E13 corresponding to Alpha_PM (20% accepted).

For the Beta variant, other than E25, E26, and E27 (corresponding to Beta_PM), 56–86% of the current traces were accepted at $\beta = 0.95$.For Beta_PM, the acceptance rate was 20-23%. For the Delta variant, we found that approximately 71–90% of the data were accepted at $\beta = 0.95$.

In the third step, all accepted current traces were low-pass filtered to mitigate high-frequency noise. Since the sampling frequency for all 39 datasets is 10 kHz, we used a low pass filter with a cutoff frequency of 3 kHz.

In the fourth step, all current traces were converted to conductance traces using the equation below:

$$Conductance\ (G_0) = \frac{Current}{V_{bias}} \tag{1}$$

where $Current$ is the measured current in nA and V_{bias} is the bias voltage in Volts. The unit G_0 in the above equation is the *conductance quantum* and is defined as follows: $G_0 = \frac{2q^2}{h} = 7.748 \cdot 10^{-5}\ S$, where q is the elementary charge and h is Planck's constant. For our datasets, the conductance values turn out to be in the range $[10^{-7.5} - 10^{-1.5}]\ G_0$.

The conductance versus time, $g(t)$, traces obtained after the previous step are then recalibrated to conductance versus distance, $g(d)$, traces using the eq. below:

$$d\ =\ 4 \cdot t \cdot Ramp\ Rate \tag{2}$$

where d is the inter-electrode distance in nm (see Fig. 2(a)) and $t = 10^{-4}$ s, corresponding to a sampling frequency of 10 kHz. The factor '4' in (2) accounts for the fact that the velocity of the top electrode is proportional to the piezo displacement, which is 4 nm/V.

C. Histogram Construction

The conductance vs. distance traces are then sampled to compute 2D and 1D histograms (viewed as probability distributions), which are used to train/test the classifiers. Fig. 2(a) shows a schematic of the experimental setup and the parameters (bias, ramp rate, current amplifier sensitivity). Fig. 2(c) and (d) show the 1D and 2D histograms constructed from a single valid conductance trace. For each of the ten COVID-19 sequences, we sample H current traces randomly and construct a conductance probability distribution (histogram). The detailed procedure for converting a current trace to a conductance trace is discussed below.

While a low value of H (ideally one) would be preferable to reduce the experimental burden, conductance histograms constructed from individual or very few current traces are extremely noisy, yielding poor classification accuracy.

Our experiments suggest that a reasonable compromise between experimental burden and accuracy is provided by choosing a baseline value of $H = 30$. In this paper, we have experimented with two different sampling methods for the conductance traces, resulting in either a *conductance histogram* or an

Authorized licensed use limited to: University of Washington Libraries. Downloaded on May 14,2025 at 23:20:13 UTC from IEEE Xplore. Restrictions apply.

average conductance histogram, and two different histogram representations, 1D or 2D. We discuss these below.

We will first clarify the distinction between a conductance histogram and an average conductance histogram. There are 5 different labels associated with Alpha_MM1: E1, E2, E3, E4, and E5. These labels correspond to different choices of experimental parameters (voltage bias, current amplifier sensitivity, and ramp rate) for the same sequence (see Table S.I of Supplemental Material). When constructing a histogram for Alpha_MM1 from H conductance traces, we have two choices: (i) sample all H traces from a particular combination of experimental parameters (e.g., E3, corresponding to a bias of 0.1 V, a current amplifier sensitivity of 10 nA/V, and a ramp rate of 10 V/s, or (ii) sample H traces randomly from all combinations of experimental parameters (mixed data sets). A histogram (properly normalized) constructed using the first sampling approach represents a conditional probability distribution of conductance values, conditioned on the specific experimental parameters. We refer to such a histogram as a *conductance histogram/probability distribution*. In contrast, a histogram constructed using the second sampling approach represents an average (unconditional) probability distribution, averaged over all combinations of experimental parameters. We refer to such a histogram as an *average conductance histogram/probability distribution*.

Irrespective of the method used to sample H conductance traces, we can construct either a 1D histogram or a 2D histogram. Since most of the inter-electrode distances are less than 0.4 nm, we divide the range of distance values, $[0 - 0.4]$ nm, into a number of distance bins, denoted by $N_{bins-Distance}$. Similarly, the range of conductance values is divided into a number of conductance bins, denoted by N_{bins}. A set of H randomly sampled conductance vs. inter-electrode distance traces is then binned along both the distance (x) axis and the conductance (y) axis, yielding a 2D conductance vs. distance histogram. By summing over the distance axis, we can marginalize the 2D histogram, yielding a 1D conductance histogram. Please refer to Fig. S1, S2, S3 and S4, S5, S6 of Supplemental Material for 1D and 2D histograms, respectively, for each experimental dataset. Conceptually, the 2D histograms can be viewed as equivalent to a time-frequency analysis (e.g., short-time Fourier Transform), whereas 1D histograms can be viewed as equivalent to a Fourier Transform. We wish to emphasize that this analogy is merely pedagogical, and we do not convert the data to the frequency domain.

D. Machine Learning Architecture

Following the procedure outlined in the previous section, we constructed 1000 1D and 2D histograms for each sequence, 70% of which are used for classifier training and validation, while the rest (30%) are used for testing.

We chose the XGBoost classifier model when the input representations are 1D histograms and a convolutional neural network (CNN) paired to an XGBoost classifier when the input representations are 2D histograms. It is important to note that the CNN and the XGBoost classifier were not trained jointly. In fact, the CNN model was trained along with a fully connected classifier, as is the norm. However, since ensemble classifiers are, in general, better suited to data that are noisy and exhibit high variance, we reckoned an ensemble classifier, operating on the features extracted by the CNN, might perform better than a single fully connected classifier. As discussed in the Results section, such an approach indeed results in an enhancement of classification accuracy by an average of 5%.

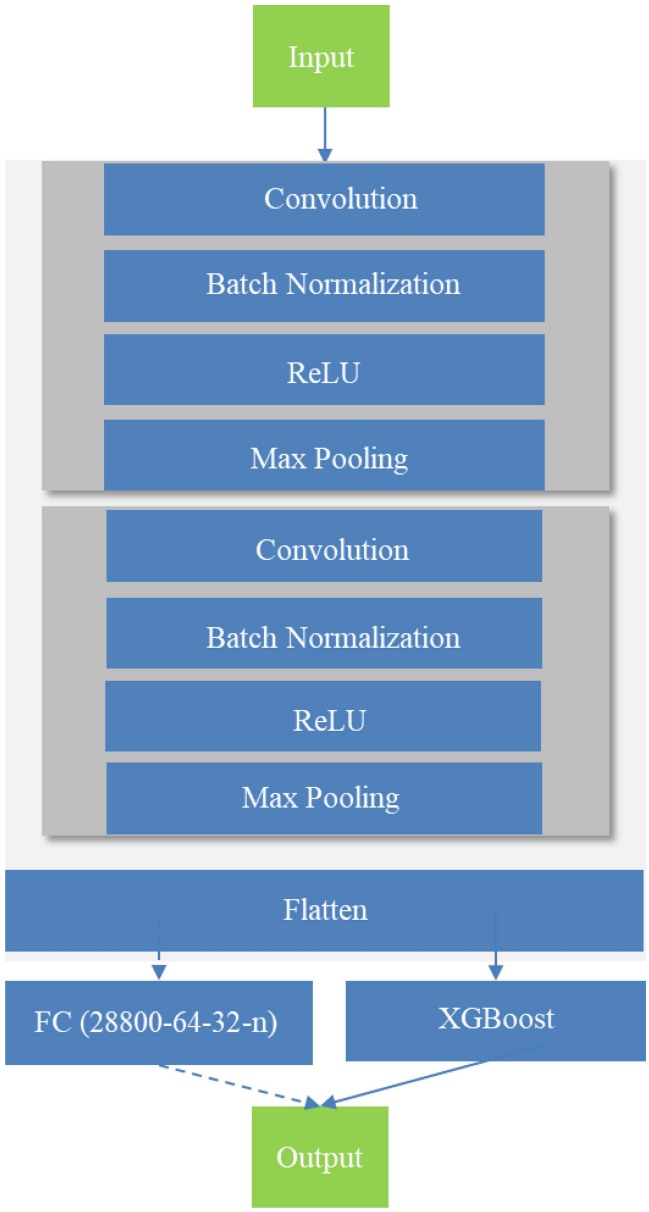


Fig. 3. The sequence of components in the stacked CNN and XGBoost model. FC stands for fully connected layer and n denotes the number of target classes.

Fig. 3 shows the architecture of our stacked CNN feature extractor and XGBoost model. We reiterate that the inputs to the XGBoost model are the outputs of the flattening layer from a CNN model trained with a fully connected classifier. For our simulations, we used the Tensorflow implementation of CNN with a learning rate based on the Adam optimizer [29] and a categorical cross-entropy loss function. All codes/data used for

Authorized licensed use limited to: University of Washington Libraries. Downloaded on May 14,2025 at 23:20:13 UTC from IEEE Xplore. Restrictions apply.

TABLE I
SUMMARY CLASSIFICATION ACCURACIES OF FOUR DIFFERENT INPUT DATA REPRESENTATIONS, ML ALGORITHMS

Approach	Histogram dimensionality	Average conductance histogram?	ML algorithm	Average accuracy	
				$H = 30$ (baseline)	$H = 10$
A1 [18]	1D	No	XGBoost	83.96%	74.50%
A2	1D	Yes	XGBoost	92.09%	82.60%
A3	2D	No	CNN + XGBoost	85.60%	75.96%
A4	2D	Yes	CNN + XGBoost	**93.27%**	**85.57%**

generating the results in this paper are available as a Python package [30].

III. RESULTS AND DISCUSSIONS

We present results from four different input data representations, which are dictated by (i) choice of 1D histograms or 2D histograms, and (ii) choice of conductance histograms or average conductance histograms. Our main motivation for studying the performance of 2D distributions is to examine whether such distributions are any more informative than 1D distributions. Table I summarizes the four possible input data representations, the ML algorithm used, and the average (over 100 trials) accuracy for a *baseline classifier*. We define a baseline classifier as one which is characterized by the following parameters: (i) prefiltering with $R^2 \leq \beta = 0.95$, (ii) histograms from $H = 30$ current traces, and (iii) number of conductance bins $N_{bins} = 600$ (for both 1D and 2D histograms) and number of distance bins $N_{bins-Distance} = 10$ for 2D histograms (equal to 1 for a 1D histogram). Average accuracy over 100 trials for sample sizes $H = 10$ and 30 are also included in Table I. In general, the performance of the classifier depends on the choice of the parameters β, H, and N_{bins}, of which the sample size parameter H has the most pronounced effect.

A. Performance of Baseline Classifier

In this section, we discuss the performance of the baseline classifier by comparing in detail the differences between approaches A1 (worst performing) and A4 (best performing) indicated in Table I. We observe that for small sample sizes such as $H = 10$, approach A4, which utilizes 2D average conductance histograms and a CNN + XGBoost classifier, always has a similar or slightly better overall accuracy than approach A2, which utilizes 1D average histograms and an XGBoost classifier. We also observe that the accuracy improves by 7-8% when average conductance histograms (whether 1D or 2D) are used, compared to conductance histograms which are conditioned on experimental parameters. This can be verified by comparing the accuracies of approaches A1 vs. A2 and approaches A3 vs. A4, shown in Table I.

Our primary research objective is as follows: "Can the PM (perfect match) and MM (mismatch) sequences for each COVID-19 variant (Alpha, Beta, Delta) be identified with high accuracy?". For each of the three variants, we treat the PM and MM sequences as different target classes. This results in three classifiers, one each for the three variants; specifically, (i) 3-class classifiers for the Alpha and Delta variants since these variants have one PM class and two MM classes, and (ii) a 4-class classifier for the Beta variant since this variant has one PM class and three MM classes.

Fig. 4 shows the confusion matrices for the Alpha, Beta, and Delta baseline classifiers; panels (a,b,c) correspond to Approach A1 [18], while panels (d,e,f) correspond to Approach A4. Please refer to Fig. S8 of Supplemental Material for the confusion matrices of all four approaches. Comparing Fig. 4(a) and (d), we see that 1D histograms work reasonably well ($\approx 90\%$ accuracy) for distinguishing the Alpha_MM2 class, but there is major confusion between the Alpha_MM1 and Alpha_PM classes (the 1D histograms for these classes are very similar). However, usage of 2D average histograms provides an across-the-board improvement in classification accuracy, approximately 8% for Alpha_MM2 and 13-15% for the Alpha_MM1 and Alpha_PM classes. The case for the Beta classifier is rather interesting. Comparing Fig. 4(b) and (e), we can see that although 1D histograms can distinguish the Beta_PM class perfectly, the accuracy is approximately 75-78% for the three mismatch classes. This suggests that 1D histograms for the Beta_PM class must be strikingly dissimilar to the three mismatch classes, which is indeed the case, as evidenced by Fig. S2 of Supplemental Material. Usage of 2D average histograms boosts the accuracies of the mismatch class Beta_MM2 by about 20% and Beta_MM3 by about 12%, although there is still room for improvement as far as Beta_MM1 and Beta_MM3 classes are concerned (roughly 15% confusion between these two classes). Finally, comparing Fig. 4(c) and (f), we observe that 2D average histograms provide an average accuracy boost of approximately 4-10% for the Delta classes.

B. Performance Analysis of Baseline Classifiers W.R.T Sample Size H

In this section, we address the impact of the parameter H, which tends to affect classifier accuracy the most. Intuitively, we expect larger values of H to yield better classifier models since the relatively high stochasticity inherent in the current

Authorized licensed use limited to: University of Washington Libraries. Downloaded on May 14,2025 at 23:20:13 UTC from IEEE Xplore. Restrictions apply.

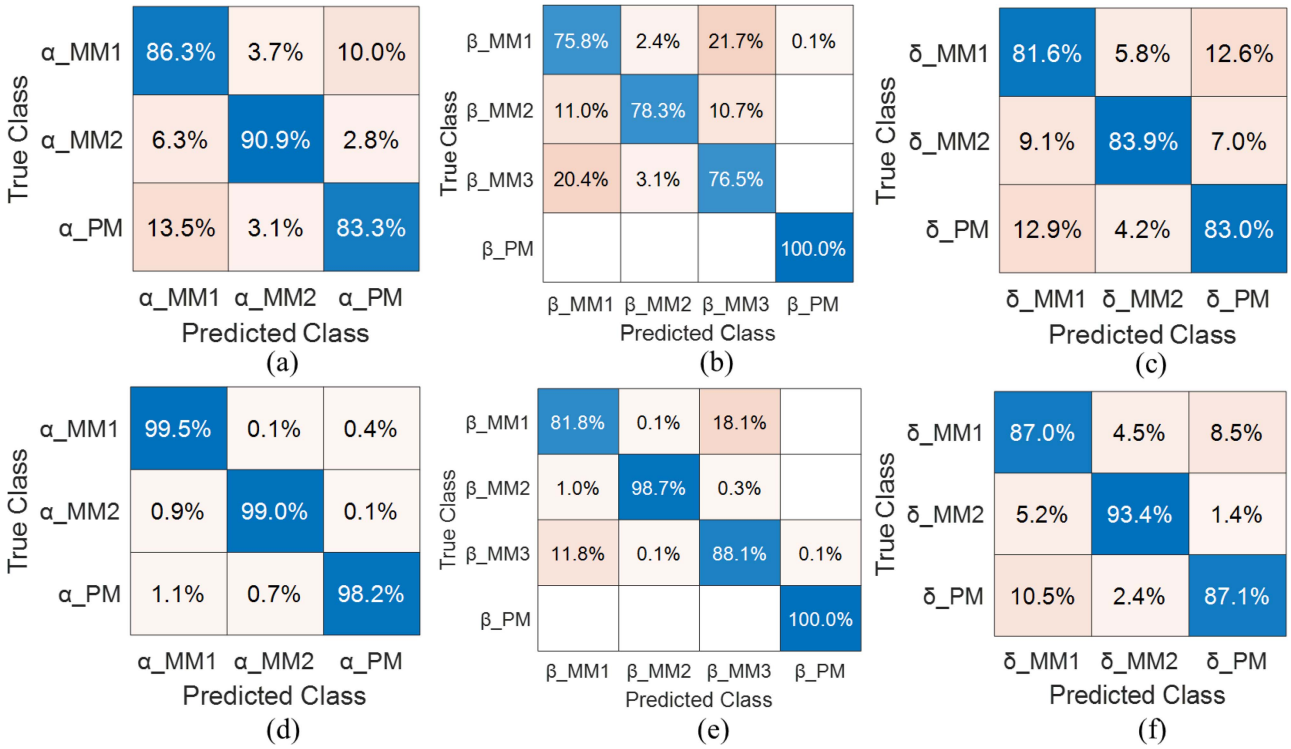


Fig. 4. Confusion matrices for baseline classifiers: (a), (b), (c) for approach A1 (d), (e), (f) for approach A4. The differences between these two approaches are summarized in Table I.

traces will tend to be 'averaged out' as H increases, revealing the underlying true probability distribution. From a 'usability' perspective, the value of $H = 1$ is ideal since it implies that a sequence determination can be made at run time after every experiment. However, the state-of-the-art experimentation currently precludes operating at an ideal value of H.

Fig. 5 shows the overall classification accuracy for all four approaches indicated in Table I with respect to the sample size H, while keeping $\beta = 0.95$, $N_{bins-Distance} = 10$, and $N_{bins} = 600$. A similar plot over an extended range of H, $H \in [10100]$, appears in Fig. S9 of Supplemental Material. The reason why we study all four approaches in this section is that Fig. 5 provides some clues regarding the relative efficacies of the two data representational attributes, (i) 2D vs. 1D histograms and (ii) average histogram or not, as discussed in the subsequent paragraphs.

From the top row in Fig. 5, we observe that for the Alpha variant, approach A4 is better than A2, which is significantly better (about 10%) than A3 and A1. This suggests that averaged histograms are more impactful in the case of Alpha variants, particularly for the Alpha PM and Alpha_MM1 sequences. We also observe that approach A4 is approximately 95% accurate for Alpha variants when $H = 10$, compared to $\approx 87\%$ for approach A2. However, as H increases, the two approaches become comparable and are visually indistinguishable for $H \geq 40$. Therefore, for the Alpha variant, we conclude that approach A4 is preferable over approach A2, particularly for smaller values of H.

From the bottom row in Fig. 5, we observe that the plots for the Delta variant are slightly different from the Alpha variant. In this case, approach A2 is the best overall, followed by approach A4, even for smaller values of H. In fact, for the Delta_MM1 sequence, A2 is approximately 6% better than A4 when $H = 10$. We also observe that neither of the two approaches can provide a 95% classification accuracy or better (our intended benchmark) for all three Delta variants even when $H = 50$. In general, we have found that Delta variants are harder to classify than Alpha variants.

Finally, the case for the Beta variant turns out to be significantly different than the Alpha and Delta variants. From the middle row in Fig. 5, we observe that the Beta_PM class is easiest to classify, reaching almost 100% accuracy at $H = 10$, irrespective of which of the four approaches is used. For the Beta_MM1 and Beta_MM3 sequences, averaged 2D histograms (approach A4) perform marginally better than non-averaged 2D histograms (approach A3), and either of these two approaches is better than approaches A1 and A2 which utilize 1D histograms. This phenomenon is similar to what we observed for the Alpha sequences. For Beta_MM2, however, approach A2 performs marginally better than approach A4, as is the case for the Delta variant sequences. In fact, if we compare the blue/green and red/black curves for Beta_MM2, we observe that averaged histograms provide almost 23-24% improvement over non-averaged histograms when $H = 10$.

We now summarize the observations/inferences from the preceding discussion. While we embarked on this analysis with

Authorized licensed use limited to: University of Washington Libraries. Downloaded on May 14,2025 at 23:20:13 UTC from IEEE Xplore. Restrictions apply.

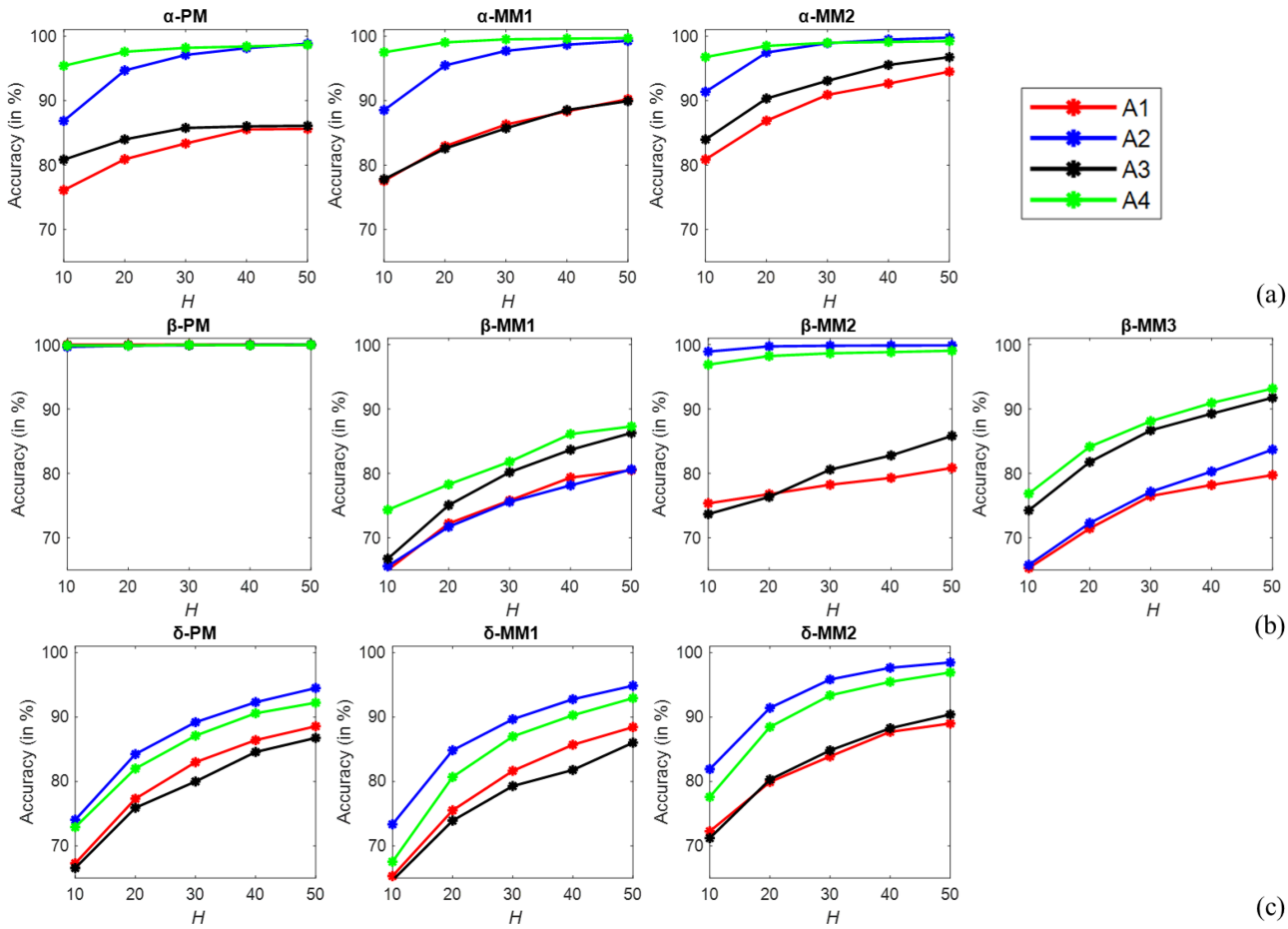


Fig. 5. Performance analysis of baseline classifiers with respect to H for the four approaches indicated in Table I. (a) Alpha sequences, (b) Beta sequences, and (c) Delta sequences. The same color scheme has been used to distinguish between the four approaches.

the goal of being able to state unequivocally which of the four approaches is best, it turns out that no straightforward answer exists. Disregarding the Beta_PM sequence for which all four approaches provide 100% accuracy even for $H = 10$, we find that averaged 2D histograms (approach A4) perform best for all three Alpha sequences and the Beta_MM1/MM3 sequences. However, averaged 1D histograms (approach A2) perform best for all three Delta sequences and the Beta_MM2 sequence. Our investigations suggest that while averaged conductance histograms are preferred over conditional histograms when experiments on COVID-19 variants and their mismatches are conducted using a range of experimental parameter values, the choice of a 1D or 2D histogram is sequence-dependent. The exact reason why some sequences benefit from a time/distance indexed conductance distribution, but not all, is not understood at this point and is subject to further theoretical investigation.

C. Impact of Applied Bias on Classifier Accuracy

So far, we have discussed the performance of our baseline classifiers and their sensitivities to the *histogram related parameter*, H. In this section, we explore how the *experimental parameter*, voltage bias, impacts classification accuracy. Obviously, this analysis necessitates a fine-grained data labeling scheme that is cognizant of the sequence and the voltage bias. Since the majority of datasets in Table S.I of Supplemental Material have been recorded using a current amplifier with a sensitivity of 10 nA/V and a ramp rate of 10 V/s, we first filtered Table S.I of Supplemental Material based on these two parameters. Subsequently, we combined datasets which pertain to the same sequence and applied bias.

For example, for the Alpha variant, the filtered datasets are E3, E4, E5, E8, E9, E14, and E15, all of which were recorded using a 10 nA/V amplifier and a 10 V/s ramp rate, although they differ in the applied bias (0.10 V, 0.15 V, and 0.20 V). We refer to this subset of seven datasets as the Alpha variant for our analysis in this section. For the Beta variant, the filtered datasets are E16, E17, E18, E21, E22, E23, E24, and E27, which differ in the applied bias (0.05 V and 0.10 V). However, since E16 and E18 correspond to the same sequence (Beta_MM1) and a bias value of 0.10 V, we combined these two datasets. Similarly, E23 and E24 were combined since they both correspond to Beta_MM3 and a bias value of 0.10 V.

This resulted in six datasets for the Beta variant, E16 ∪ E18, E17, E21, E22, E23 ∪ E24, and E27, where ∪ denotes the set union. Following an identical procedure, we end up with eight

Authorized licensed use limited to: University of Washington Libraries. Downloaded on May 14,2025 at 23:20:13 UTC from IEEE Xplore. Restrictions apply.

TABLE II
CLASSIFIER ACCURACIES FOR ALPHA, BETA, AND DELTA VARIANTS AS A FUNCTION OF APPLIED BIAS

		Dataset	Accuracy	Dataset	Accuracy	Dataset	Accuracy
Voltage Bias		0.10 V		0.15 V		0.20 V	
Alpha variant	MM1	E3	99.47 %	E4	86.30 %	E5	74.92 %
	MM2	E8	99.62 %	E9	99.24 %		
	PM	E14	99.78 %			E15	37.10 %
Mean accuracy			99.62 %		97.77 %		56.01 %

Voltage Bias		0.05 V		0.10 V	
Beta variant	MM1	E17	99.30 %	E16 ∪ E18	90.41 %
	MM2			E21	84.12 %
	MM3	E22	99.88 %	E23 ∪ E24	92.61 %
	PM			E27	97.89 %
Mean accuracy			99.59 %		91.26 %

Voltage Bias		0.03 V		0.10 V		0.20 V	
Delta variant	MM1	E51 ∪ E52	95.45 %	E53 ∪ E54	89.83 %		
	MM2	E55	99.92 %	E56 ∪ E57	90.41 %	E58	99.99 %
	PM	E59 ∪ E60	98.89 %	E61	95.36 %	E62	99.69 %
Mean accuracy			98.09 %		91.87 %		99.84 %

All histograms generated using $H = 100$. All datasets referenced in this table were recorded using a current amplifier of sensitivity $10\ nA/V$ and a ramp rate of $10\ V/s$.

datasets for the Delta variant, E51 ∪ E52, E53 ∪ E54, E55, E56 ∪ E57, E58, E59 ∪ E60, E61, and E62.

Next, we generated histograms using approach A3 (2D non-averaged histograms, see Table I) and $H = 100$. A reasonably large value of H was chosen to minimize the effect of sampling-induced uncertainties on classifier performance. It is important to note that each such histogram represents a *unique* combination of the tuple (sequence, voltage bias). These histograms described in the previous paragraph are then used to train and test independent classifiers for each of the three variants (Alpha, Beta, and Delta): (i) a 7-class classifier for the Alpha variant, (ii) a 6-class classifier for the Beta variant, and (iii) an 8-class classifier for the Delta variant. Table II shows the classifier accuracies for Alpha, Beta, and Delta variants as a function of applied bias. The confusion matrices for these three classifiers are shown in Fig. S10, S11, and S12 of Supplemental Material. Results for classifiers trained on histograms with $H = 30$ are shown in Table S.II of Supplemental Material.

A general observation from Table II is that lower biases result in higher accuracies if we restrict ourselves to a moderate bias range $[0.03,\ 0.15]$ V. This is true for all three variants. However, when the bias is 0.20 V, we observe contradictory behaviors for the Alpha and Delta variants. While the accuracy drops precipitously for the Alpha sequences E5 (Alpha_MM1) and E15 (Alpha_PM) at a bias of 0.20 V, it approaches almost 100% for the Delta variant sequences E58 (Delta_MM2) and E62 (Delta_PM). This contradictory behavior can be validated by using an information-theoretic criterion, namely the Jensen-Shannon metric, as explained in Section 5 of Supplemental material (see Fig. S13 in particular).

The preceding discussion regarding the contradictory behavior exhibited by the Alpha and Delta variant sequences at a bias of 0.20 V raises an interesting question: are some DNA sequences naturally more susceptible to higher applied bias? Could it be due to some structural deformations within the sequence, which are induced by higher bias values? We leave this as an open question for now, pending additional theoretical investigation.

IV. CONCLUSION

The identification of genetic material and chemical detection from the conductance of single molecules has been pursued for over a decade. The noisy nature of the experimental measurements on single molecules has been a challenge in enabling robust identification. Further, physics-based modeling has failed to theoretically discern between molecules by accounting for the possible contact configurations and environmental fluctuations. In this paper, we propose a methodology based on multiple machine learning techniques and two different methods for data representation to identify DNA strands with high accuracy. Accompanying the developed method is the release of a software package that can be used for a wide range of molecular conductance data [30].

Experimentally, the distance between the metal electrodes is varied, and conductance is measured as a function of time. As a result of the intrinsically noisy nature of the experiments, the

Authorized licensed use limited to: University of Washington Libraries. Downloaded on May 14,2025 at 23:20:13 UTC from IEEE Xplore. Restrictions apply.

conductance values are not completely reproducible if experiments are repeated on the same molecule. This prompted us to represent the data as conventional 1D conductance probability distributions and, additionally, as a 2D probability distribution, which depends on both the conductance value and inter-electrode separation. We studied the accuracy of detection both with pure datasets (no averaging over experimental parameters) and mixed datasets that involved averaging the conductance distribution over experimental parameters. In conjunction with 2D distributions, we have proposed an architecture consisting of a convolutional neural network paired with an XGBoost classifier. Our work shows that (i) averaged (mixed data) distributions have a significant impact on classifier accuracy, boosting it by as much as 21%, (ii) 2D conductance distributions are beneficial for some sequences, and (iii) the backend XGBoost classifier whose input is feature maps from a convolutional neural network provides better accuracy than a fully connected layer usually adopted with a convolutional neural network. Finally, the classifier models were used to study the impact of the applied voltage between the two electrodes on accurate sequence detection. We find evidence that relatively large biases can be detrimental to the development of accurate classifier models. Overall, our classification approach exhibits significant potential for accurate, amplification-free DNA sequence identification. Importantly, while our analysis focused on COVID-19 data, the computational method proposed in this paper should be of use in analyzing arbitrary single-molecule conductance data.

Acknowledgment

The authors acknowledge Zahra Aminiranjbar and Josh Hihath for discussion of the data generated.

References

[1] J. Ritchie, "Probabilistic DNA evidence: The laypersons interpretation," *Australian J. Forensic Sci.*, vol. 47, no. 4, pp. 440–449, Oct. 2015, doi: 10.1080/00450618.2014.992472.

[2] F. E. Dewey, S. Pan, M. T. Wheeler, S. R. Quake, and E. A. Ashley, "DNA sequencing clinical applications of new DNA sequencing technologies," *Circulation*, vol. 125, no. 7, pp. 931–944, 2012, doi: 10.1161/CIRCULATIONAHA.110.972828.

[3] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, Mar. 2013, doi: 10.1126/science.1235122.

[4] Y. Li et al., "Detection and identification of genetic material via single-molecule conductance," *Nat. Nanotechnol.*, vol. 13, no. 12, pp. 1167–1173, Dec. 2018, doi: 10.1038/s41565-018-0285-x.

[5] K. G. G. Pattiya Arachchillage, S. Chandra, A. Piso, T. Qattan, and J. M. Artes Vivancos, "RNA BioMolecular electronics: Towards new tools for biophysics and biomedicine," *J. Mater. Chem. B*, vol. 9, no. 35, pp. 6994–7006, 2021, doi: 10.1039/D1TB01141C.

[6] B. Xu, P. Zhang, X. Li, and N. Tao, "Direct conductance measurement of single DNA molecules in aqueous solution," *Nano Lett.*, vol. 4, no. 6, pp. 1105–1108, Jun. 2004, doi: 10.1021/nl0494295.

[7] T. Nishino and P. T. Bui, "Direct electrical single-molecule detection of DNA through electron transfer induced by hybridization," *Chem. Commun.*, vol. 49, no. 33, 2013, Art. no. 3437, doi: 10.1039/c3cc38992h.

[8] K. G. Gunasinghe Pattiya Arachchillage et al., "A single-molecule RNA electrical biosensor for COVID-19," *Biosens Bioelectron*, vol. 239, Nov. 2023, Art. no. 115624, doi: 10.1016/j.bios.2023.115624.

[9] E. M. Dief and N. Darwish, "SARS-CoV-2 spike proteins react with Au and Si, are electrically conductive and denature at 3×10^{8} V m^{-1} : A surface bonding and a single-protein circuit study," *Chem. Sci.*, vol. 14, no. 13, pp. 3428–3440, 2023, doi: 10.1039/D2SC06492H.

[10] E. M. Dief, P. J. Low, I. Díez-Pérez, and N. Darwish, "Advances in single-molecule junctions as tools for chemical and biochemical analysis," *Nat. Chem.*, vol. 15, no. 5, pp. 600–614, May 2023, doi: 10.1038/s41557-023-01178-1.

[11] B. Zhang et al., "Role of contacts in long-range protein conductance," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 13, pp. 5886–5891, Mar. 2019, doi: 10.1073/pnas.1819674116.

[12] L. Xiang et al., "Conductance and configuration of molecular gold-water-gold junctions under electric fields," *Matter*, vol. 3, no. 1, pp. 166–179, Jul. 2020, doi: 10.1016/j.matt.2020.03.023.

[13] L. E. Scullion, E. Leary, S. J. Higgins, and R. J. Nichols, "Single-molecule conductance determinations on $HS(CH_2)_4O(CH_2)_4SH$ and $HS(CH_2)_2O(CH_2)_2O(CH_2)_2SH$, and comparison with alkanedithiols of the same length," *J. Phys.: Condens. Matter*, vol. 24, no. 16, Apr. 2012, Art. no. 164211, doi: 10.1088/0953-8984/24/16/164211.

[14] T. Nishino, H. Shiigi, M. Kiguchi, and T. Nagaoka, "Specific single-molecule detection of glucose in a supramolecularly designed tunnel junction," *Chem. Commun.*, vol. 53, no. 37, pp. 5212–5215, 2017, doi: 10.1039/C6CC09932G.

[15] H. Li et al., "Large variations in the single-molecule conductance of cyclic and bicyclic silanes," *J. Amer. Chem. Soc.*, vol. 140, no. 44, pp. 15080–15088, Nov. 2018, doi: 10.1021/jacs.8b10296.

[16] H. Chen et al., "Promotion and suppression of single-molecule conductance by quantum interference in macrocyclic circuits," *Matter*, vol. 4, no. 11, pp. 3662–3676, Nov. 2021, doi: 10.1016/j.matt.2021.08.016.

[17] Z. Wang et al., "Electrochemically controlled rectification in symmetric single-molecule junctions," *Proc. Nat. Acad. Sci. USA*, vol. 119, no. 39, Sep. 2022, Art. no. e2122183119, doi: 10.1073/pnas.2122183119.

[18] Y. Wang, M. Alangari, J. Hihath, A. K. Das, and M. P. Anantram, "A machine learning approach for accurate and real-time DNA sequence identification," *BMC Genomic.*, vol. 22, no. 1, Dec. 2021, Art. no. 525, doi: 10.1186/s12864-021-07841-6.

[19] Y. Komoto, J. Ryu, and M. Taniguchi, "Machine learning and analytical methods for single-molecule conductance measurements," *Chem. Commun.*, vol. 59, no. 45, pp. 6796–6810, Jun. 2023, doi: 10.1039/D3CC01570J.

[20] S. Chang et al., "Chemical recognition and binding kinetics in a functionalized tunnel junction," *Nanotechnology*, vol. 23, no. 23, Jun. 2012, Art. no. 235101, doi: 10.1088/0957-4484/23/23/235101.

[21] T. Fu, Y. Zang, Q. Zou, C. Nuckolls, and L. Venkataraman, "Using deep learning to identify molecular junction characteristics," *Nano Lett.*, vol. 20, no. 5, pp. 3320–3325, May 2020, doi: 10.1021/acs.nanolett.0c00198.

[22] J. Ryu, Y. Komoto, T. Ohshiro, and M. Taniguchi, "Single-molecule classification of aspartic acid and leucine by molecular recognition through hydrogen bonding and time-series analysis," *Chem.: Asian J.*, vol. 17, no. 13, Jul. 2022, Art. no. e202200179, doi: 10.1002/asia.202200179.

[23] D. Stefani et al., "Conformation-dependent charge transport through short peptides," *Nanoscale*, vol. 13, no. 5, pp. 3002–3009, 2021, doi: 10.1039/D0NR08556A.

[24] S. Tao et al., "Revealing conductance variation of molecular junctions based on an unsupervised data analysis approach," *Electrochim Acta*, vol. 449, May 2023, Art. no. 142225, doi: 10.1016/j.electacta.2023.142225.

[25] D. Lin et al., "Using weakly supervised deep learning to classify and segment single-molecule break-junction conductance traces," *ChemPhysChem*, vol. 22, no. 20, pp. 2107–2114, Oct. 2021, doi: 10.1002/cphc.202100414.

[26] N. D. Bamberger, J. A. Ivie, K. N. Parida, D. V. McGrath, and O. L. A. Monti, "Unsupervised segmentation-based machine learning as an advanced analysis tool for single molecule break junction data," *J. Phys. Chem. C*, vol. 124, no. 33, pp. 18302–18315, Aug. 2020, doi: 10.1021/acs.jpcc.0c03612.

[27] Z. Aminiranjbar et al., "Identifying SARS-CoV-2 Variants using single-molecule conductance measurements," *ACS Sens*, vol. 9, pp. 2888–2896, May 2024, doi: 10.1021/acssensors.3c02734.

[28] "Tracking SARS-CoV-2 variants," 2022. Accessed: Apr. 06, 2022. [Online]. Available: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[30] Y. Wang, H. Wang, A. K. Das, and M. P. Anantram, "SMBJClassifier," Accessed: Feb. 28, 2024. [Online]. Available: https://github.com/ethanwyr/SMBJClassifier, 2024. [Online]. Available: https://github.com/ethanwyr/SMBJClassifier

Authorized licensed use limited to: University of Washington Libraries. Downloaded on May 14,2025 at 23:20:13 UTC from IEEE Xplore. Restrictions apply.



Yiren Wang received the BS degree in electrical engineering, the MS degree in electrical engineering, and the MS degree in applied mathematics from the University of Washington, USA, in 2019, 2021, and 2024. He is currently working toward the PhD degree in electrical and computer engineering with the University of Washington and anticipates graduating, in 2025. From 2019, he has been part of M.P. Anantram's group modeling charge transport in nanoscale devices. His research interests include nanotechnology, bioinformatics, modeling charge transport in DNA, quantum mechanical modeling of nanodevices, the electrical properties of DNA/RNA molecules, statistical data analysis, and machine learning.



Hongning Wang participates in the internship program with the Department of Electrical and Computer Engineering, University of Washington, USA. He is a senior with Eastlake High School, Sammamish, WA, USA. He will matriculate to university in Sep. 2025 and plans to study computer science and applied mathematics. His research interests include machine learning and computer vision.



Arindam K. Das received the MS and PhD degrees in electrical engineering from the University of Washington, USA, in 1999 and 2003, respectively. From 2003 to 2007, he was a postdoctoral research associate jointly with the Department of Electrical Engineering, Department of Aeronautics and Astronautics, University of Washington. From 2007 to 2012, he was a research scientist with the Applied Physics Laboratory, University of Washington. He is currently an associate professor with the Computer Science and Electrical Engineering Department, Eastern Washington University, USA, and an Affiliate associate professor with the Electrical Engineering Department, University of Washington. He is also an independent consultant to several technology and biotech companies. His current research interests include broadly in the areas of optimization theory and applications, machine learning, and metaheuristic optimization.



M. P. Anantram (Senior Member, IEEE) received the BSc degree in applied science from P.S.G. College of Technology, Coimbatore, India, the MSc degree in physics from the University of Pune, India, and the PhD degree in electrical engineering from Purdue University, USA. He has worked at the NASA Ames Research Center's Center for Nanotechnology and was a professor with the University of Waterloo, Canada before joining the University of Washington, USA. His group at the University of Washington works on the theory, algorithm and application of modeling methods for nanoscale materials and devices. The group's current focus is on modeling of electron transport in DNA and peptide nanostructures, and multi-scale modeling of memory devices.

Authorized licensed use limited to: University of Washington Libraries. Downloaded on May 14,2025 at 23:20:13 UTC from IEEE Xplore. Restrictions apply.