Examination of DNA Electronic Properties Utilizing Machine Learning, Statistical, and Modeling Methods

Yiren Wang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Anant M.P. Anantram, Chair

Arindam Kumar Das

Shih-Chieh Hsu

Program Authorized to Offer Degree:

Electrical and Computer Engineering

University of Washington

Abstract

Examination of DNA Electronic Properties Utilizing Machine Learning, Statistical, and Modeling Methods

Yiren Wang

Chair of the Supervisory Committee:

Anant M.P. Anantram

Department of Electrical and Computer Engineering

This thesis focuses on the examination of DNA electronic properties through the utilization of statistical feature extraction, machine learning algorithms, density functional theory (DFT), and Green's function method. The common discrepancies observed between experimental and theoretical results have driven us to develop a comprehensive theoretical simulation and modeling approach that is based on experimental outcomes. Initially, we investigate the effect of counterions and solvent dielectric on the conductance of B-form DNA, aiming to comprehend how they modulate DNA electronic properties. Our simulation results indicate that in dry DNA, the presence of counterions affects electron transmission at the lowest unoccupied molecular orbital (LUMO) energies. However, in a solution, counterions play a negligible role in transmission. By employing the polarizable continuum model calculations, we demonstrate that

the transmission is significantly higher at both the highest occupied (HOMO) and lowest unoccupied molecular orbital (LUMO) energies in a water environment as opposed to a dry environment.

Subsequently, we present a DNA sequence identification system based on the conductance characteristics of short-strand sequences, including a group that differs by a single mismatch. By training a gradient-boosted tree classifier model (XGBoost) on 1D conductance histograms, we achieve remarkably high accuracy, ranging from approximately 96% for sequences with a single mismatch to 99.5% otherwise. These accuracy metrics are attained in near real-time with a minimal number of ten SMBJ (Single Molecule Break Junction) measurements rather than hundreds or thousands. To improve the robustness of the identification system, by targeting specific sequences with a single base mismatch, we propose an approach based on combining XGBoost and a convolutional neural network with different input feature representations: 2D conductance probability distributions, with averaging over the experimental parameters. While the adoption of a 2D probability distribution is helpful with respect to classifier accuracy, we find that averaged conductance probability distributions are much more impactful and significantly enhance the prediction accuracy. Our quantitative analysis of multiple sequences shows an impressive performance boost (approximately 10%) for all sequences. Another key result that emerges from the method developed is evidence that lower voltage bias values produce more accurate classification accuracy.

Moreover, to further address the low signal-to-noise ratio inherent in SMBJ measurements, we developed a Piecewise Linear Approximation method, which extracts plateau segments from a noisy time series trace. This method could remove noise content more effectively and keep all useful conductance information unchanged. The resulting conductance histograms, which are

constructed purely from plateau segments, provide huge benefits in training subsequence machine learning models and focus on learning essential conductance features. Even with a sample size as small as five, our classification system is able to maintain its high accuracy. The enhanced machine learning approach with the Piecewise Linear Approximation method could significantly improve the efficiency and accuracy of SMBJ methods, making them more viable for practical applications. Although the basis of the analysis in this thesis is time series conductance data of less than a hundred DNA/RNA strands and their single base mismatches, our method is generally applicable to other single-molecule electrical data. We posit that our approach represents an emerging alternative to existing DNA sequence identification methods and should be of use in analyzing single molecule conductance data for sequence identification.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my highest gratitude to my advisor Prof. Anant M. P. Anantram for his continuous support, encouragement, and guidance during my undergraduate and Ph.D. study. I would like to thank my secondary advisor Prof. Arindam Kumar Das for his active engagement and patient supervision in enabling my significant achievement throughout my Ph.D. study. Special thanks to my committee members - Prof. Shih-Chieh Hsu from Department of Physics, Prof. Krithika Manohar from Department of Mechanical Engineering, and Prof. Sreeram Kannan from Department of Electrical and Computer Engineering.

I would like to thank our international collaborators Prof. Ersin Emre Oren and his fantastic group at TOBB University of Economics and Technology (in Ankara, Turkey), especially Dr. Busra Demir. I would also like to thank our experimental collaborators Prof. Josh Hihath and his excellent group at Arizona State University, especially Dr. Zahra Aminiranjbar, Dr. Bo Liu, and Dr. Mashari Alangari. My gratitude goes to my colleagues for all their support on both work and life: Dr. Hashem Mohammad, William Livernois, Arpan De, Olayian Alolayian, Shiang-Bing Chiu, Simon Cao, Paritosh Singh. I also want to thank two high school students I have worked with Vikram Khandelwal and Hongning Wang.

Last but not least, I want to thank my family for being supportive for so many years. I am grateful to my parents for their sacrifices, dedication, and support, which made it possible for me to study abroad for more than 10 years. I want to nominate my grandparents as my lifelong role models, whose values and dedications continue to inspire me. I would like to thank my dog Bagel and my hamster Taro. Finally, I want to express my love to my partner Zi Ye. Without her I cannot go through all the hard times and complete my Ph.D.

1 Introduction

DNA (deoxyribonucleic acid) is the most essential and fundamental macromolecule that underlies the functioning of all life forms on our planet. Thorough research on DNA molecules has propelled them to become one of the leading materials in biomolecular electronics and nanotechnology primarily due to their molecular recognition, self-assembly, and long-range charge transport properties [1]. With the rapid advancements in nanoelectronics, traditional silicon-based devices gradually become unable to fulfill the continuous demands of decreasing component size and reducing energy consumption. Conversely, biomolecular electronic devices offer a potential solution by leveraging the unique properties of DNA molecules. They provide a bottom-up approach for constructing nanoscale devices with high resolution, while simultaneously ensuring energy efficiency through the utilization of chemical reactions plus enzymes [2]. Therefore, there have been extensive studies into the electronic properties of DNA, with both theoretical and experimental work, over the last two decades [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]. However, a rich diversity in conclusions has emerged depending on different experimental methods, which often leads to significant discrepancies from the theoretical simulations. This calls for the need for a comprehensive and systematic examination that combines both experimental and theoretical perspectives on charge transport through nucleic acids. In this thesis, my research focuses on investigating the influences of environmental factors on DNA charge transport, analyzing the electronic properties of DNA molecules using conductance data measured with SMBJ (Single Molecule Break Junction) experiments, designing machine learning approaches for DNA/RNA sequence identification system, and developing statistical methods for essential conductance feature extraction for all SMBJ measurements.

1.1 DNA Structure and History of DNA Sequencing

Although every living being on the earth has its unique DNA strands, all the DNA molecules are composed of the same components: sugar-phosphate backbones and nitrogenous base-pairs. And there are only four different nitrogenous bases: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G), which is called the Watson-Crick Model of DNA. Two nitrogenous bases combine to form one base pair: A and T bases will pair together; C and G bases will pair together [19]. DNA molecules exist as a nanoscale double helix structure, which stores an enormous amount of genetic information inside a tiny space. For example, every human cell has approximately 3 billion base pairs [20]. If people want to decrypt and understand all the information hiding inside the DNA strand, the first step must be finding the sequence of it. Therefore, the method to determine the arrangement of nitrogenous bases, DNA sequencing, has become one of the most indispensable techniques in our modern life. People are always trying to find quicker, cheaper, easier, and more accurate methods of DNA sequencing. In the meantime, people discovered more and more important applications to benefit our life, such as evolutionary

biology, virology (COVID-19), medicine (gene mutation diseases), forensics (find criminals), and so on [21], [22].

Since the 1970s, researchers have dedicated enormous amounts of effort and time to increase the throughput and accuracy, while also aiming to reduce the processing time and cost of DNA sequencing. Between 1970 and 2000, various fundamental methods were developed for small-volume DNA sequencing. The first major breakthrough in sequencing technology was made by Fredrick Sanger in 1977, which relied on DNA chain termination [22]. This technique, also known as "Sanger Sequencing," is one of the most widely used techniques during this period [22]. However, Sanger sequencing had limitations, as it can only sequence around 500 to 600 base pairs in a single run. The Human Genome Project, for instance, took 13 years and nearly three billion dollars to map the complete human genome [20]. Around 2008, a new generation of sequencing methods emerged and quickly replaced the traditional Sanger-based techniques in the market. These methods, commonly referred to as next-generation or short-read sequencing methods, offered substantial improvements, particularly in terms of high throughput. They enabled the parallel sequencing of thousands to millions of DNA strands simultaneously [23]. Some of the well-known commercially available techniques are 454 pyrosequencing, Illumina (Solexa) sequencing, SOLiD sequencing, and ion torrent semiconductor sequencing. Each of these methods has its own advantages, such as higher accuracy (over 99.9%), lower cost, and faster processing time [24]. Nowadays, owing to the remarkable developments in high-throughput sequencing techniques, it is possible to map the entire human genome within a day at a cost of approximately $1000.

However, short-read sequencing methods share some common drawbacks. They require amplified identical DNA strands and the fragmentation of long strands into smaller segments (around 300 base pairs) to avoid the error rate increasing exponentially. Therefore, researchers have shifted their focus towards a different group of high-throughput sequencing methods known as long-read or third-generation sequencing methods [25]. These methods are designed to sequence one entire DNA molecule at once, which can span more than 30,000 base pairs. Long-read methods have been found to be particularly advantageous in sequence assembly and resolving complex regions of DNA strands [26]. The two dominant long-read sequencing methods are single-molecule real-time (SMRT) sequencing by Pacific Bioscience and nanopore DNA sequencing by Oxford Nanopore Technologies. However, the error rate of these long-read methods is still higher than short-read methods, where the SMRT sequencer claims an error rate of less than 1% and the nanopore sequencer claims an error rate of less than 5%. Furthermore, the processing time and general cost of long-read methods are also less favorable compared to short-read methods [25]. Therefore, these huge challenges must be overcome before long-read methods can become reliable and replace short-read methods commercially.

About 20 years ago, researchers conducted the first measurements of current flow through DNA molecules, which subsequently prompted numerous experiments and theoretical modeling

efforts to determine the electrical properties of DNA molecules and DNA nucleotides [2]. Researchers discovered that many features could affect the electrical properties of DNA, including the base sequence, strand length, DNA structure, and environmental conditions [27]. As a result, experiments have shown contradictory results regarding the electrical properties of DNA, with DNA being described concurrently as an insulator, conductor, and semiconductor. Researchers also found that each of the four nitrogenous bases (A, T, C, G) possess distinct electrical properties. For instance, the Guanine base has the lowest oxidation potential [28]. However, despite these experimental results, a well-established theoretical model that could explain the principles behind DNA electron transportation and the differences among electrical behavior of the nitrogenous bases is currently lacking [27]. The identification of DNA molecules and nucleotides using nanoelectronics remains an immensely critical aspect, particularly for the advancement of third-generation sequencing methods. By leveraging the advantage of NOT requiring synthetic DNA amplification, nanoelectronics based methods have the potential to hugely reduce laboratory costs, DNA sample preparation time, chemical reaction time, and associated potential errors.

1.2 Solvent Conditions of DNA Molecules

DNA is one of the leading materials in molecular electronics due to its self-assembly property and long-range charge transport [1]. It naturally exists in a solvent environment, surrounded by water and salt molecules. Depending on the sequence of DNA, the environmental factors, such as the dielectric constant of solvents and the position and/or local density of counterions surrounding DNA, become essential in determining DNA conformation and thus conductance. The electronic properties of DNA have been actively studied over the last two decades [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]. However, the complex environment that a DNA molecule is in makes it challenging to analyze and interpret experimental results. Thus, modeling methods of varying degrees of complexity and accuracy become vital to understanding features in DNA conductance.

In several experiments, DNA conductance has been measured in both hydrated and dehydrated conditions [9], [29]. These studies cannot easily separate the solvent contribution to conductance as some solvent molecules remain attached to the DNA even in the dehydrated case [30], [31], [32]. Initially, theoretical studies of charge transport were performed in a dehydrated environment [33] due to high computational costs. More recently, studies have begun modeling the role of the solvent environment in charge transport [1], [34], [35]. Kubař et al. suggested that the on-site energies can change with time by about 0.4 eV due to fluctuations in the solvent [36]. The off-diagonal hopping parameters can, in turn, affect DNA structure and transmission. Further, some studies included the influence of conformation and the solvent to investigate their joint role in determining conductance [27], [36]. Reference [27] suggested that solvent effects can decrease the sequence dependence of conductance. On the other hand, recent papers [16], [37] show that even a single base-pair mismatch can be detected experimentally in the presence

of a solvent. Further, [38] analyzed the experimentally obtained values of conductance using a pure machine learning approach and concluded that the experimental data show differences in the conductance of strands with a single mismatch.

Wolter et al. used ab initio molecular dynamics simulation and concluded that DNA in a micro-hydrated environment, which retains the structure of the close solvation shell, shows charge transport properties similar to fully solvated DNA. They disagreed with experimental studies suggesting a strong humidity dependence of DNA conductance by arguing that in the presence of high humidity or solvent content, the conductivity of the residual solvent (not DNA) is measured [34]. The early work of Berashevich et al. studied the influence of humidity on DNA by hydrating base pairs with water molecules attached to the DNA bases [39]. They found hydration increased the band gap of bases and concluded that this would make the conductance sensitive to the water environment.

The effect of solvent on the structural properties of DNA has also been studied computationally using molecular dynamics (MD) simulations [33], [40], [41], [42], [43]. These studies show that the interaction between DNA molecules and the solvent contributes to fluctuations, leading to changes in the energy of both molecular orbitals and ionization potentials. Previously, it also has been demonstrated that the solvent environment affects DNA conformation [15], [44], [45], [46]. The common B form is found at neutral pH and normal saline [47], while the A form prefers dry or dehydrated conditions. On the other hand, the A form has a shorter distance between the base pairs when compared to the B form. The base pairs of A form DNA are located away from the helical axis and are closer to the major groove [47]. Based on the prior work summarized above, it is challenging to distinguish the role of solvent environment and conformation from one another in modeling studies. To only reveal the role of the solvent, in this model study, we kept the DNA structure fixed and only changed the dielectric constant of the environment in our calculations. Apart from the solvent, counterions are also present close to the sugar-phosphate backbone. Positively charged counterions (such as Na+) are attracted to the negative charge on the phosphate group of the backbone. Prior studies have considered the effect of the counterions and concluded that they also affect DNA conductance. However, it has been difficult to reach clear conclusions due to the plethora of techniques and approximations [1], [27], [34], [48], [49].

1.3 Machine Learning Approaches for DNA Sequence Identification

Currently, no molecular detection and identification technique exists that can speedily, reliably, and without the need for amplification steps, detect or identify a low copy or a wide range of single biologically relevant molecules. A promising technique is an all-electronic method that would identify DNA/RNA based on the characteristic measurements of current [2]. This all-electronic DNA sequence identification system experiences nonlinear interactions between the substrate, sample, environment, and measuring system that are inherently stochastic [27], [28]. It has been extremely challenging for physics-based models to capture the differences

in current between nominally different DNA strands. With the emergence of machine learning tools over the last decade, there is an increased interest in utilizing these methods to identify molecules from the current's signature. Recently, surface-enhanced Raman spectroscopy [50], scanning tunneling spectroscopy (STS) [18], [28], [51], scanning tunneling microscopy break junction (STM-BJ) [17], and the measurement of ionic current blockade in nanopores [52] have been used to identify/sequence single-stranded DNA, RNA, and peptides. Identification of small molecules from a mixture was also recently demonstrated using machine learning classification methods [37], [53], [54].

With the emerging development of machine learning, many researchers intend to utilize machine learning with DNA sequencing, especially for the development of new long-read sequencing methods. Korshoj et al. used scanning tunneling spectroscopy (STS) to measure the biophysical characters and calculate 9 different biophysical parameters of all four DNA nucleotides on one single molecule, which include LUMO, HOMO, bandgap, transition voltage, tunneling barrier height, ratio of effective mass at positive and negative biases and so on. By selecting and tuning these parameters for each nucleotide separately, they trained their machine-learning algorithm to achieve an average confidence of 80.9% in predicting DNA nucleobases [28]. The same group used a similar approach, quantum tunneling spectroscopy (QTS) to measure and calculate 12 biophysical parameters of all four RNA (A, U, C, G) nucleotides on one single molecule. After optimal parameter tuning, they achieved the highest 99.8% prediction accuracy with their new machine learning algorithm [18]. Both algorithms need an enormous amount of data to preprocess and hand-picking features to reach high accuracy. Kolmogorov et al. designed an algorithm to identify single-molecule proteins with sub-nanopores. With charge linearized translocation of protein through pores with sub-nanometer diameters, they used several machine learning methods and reached high matches with the PrNMs protein database [52].

Besides finding biophysical parameters from STS and measuring ion charge from nanopores, the conductance of DNA is also a commonly used feature. Among all the electronic properties of DNA molecules, conductance is one of the most easily obtainable information. Fu et al. used scanning tunneling microscope-based break junction (STM-BJ) to measure the conductance traces of DNA molecules. Without any further, preprocess, they used a convolutional neural network (CNN) to extract features and classify the traces, which achieved the highest 97.6%, average of 93.18%, classification accuracy [37]. They were only focused on binary classification, comparing two DNA molecules, or identifying the mixture of them. Cabosart et al. group used mechanically controlled break-junction (MCBJ) measurements to classify DNA conductance value. They transformed raw conductance traces into individual 2D histograms. With an unsupervised machine learning algorithm, K-means++, 11 different DNA datasets cluster into three conductance value classes, which significantly improved the ability to extract DNA conductance value [53]. However, there is no quantitative measurement of how accurately the clustering was performed in this particular method. Hamill et al. used a single molecular break

junction (SMBJ) to obtain conductance traces (small molecules not DNA) and convert them into individual 1D histograms. With the principal component analysis (PCA) method, they trained binary classifiers from separately measured conductance traces of two molecules. The classifier sorted a mixture of data containing both molecules into two classes with an accuracy ratio of concentration [54]. However, they cannot guarantee that the conductance traces classified into class 1 truly belonged to class 1, due to the absence of true labels for the mixture data.

Additionally, several researchers have explored/proposed single molecule all-electronic methods for DNA/RNA sequence [16], [55], [56], [57], [58], [59] and chemical [60], [61], [62], [63], [64], [65], [66], [67] 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 that can produce highly accurate results in near real-time [38], [68]. 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 [37], [69], [70], [71], [72], [73], [74]. 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.

1.4 Optimal Piecewise-Linear Function Approximation Method

The relentless pursuit of miniaturization in electronic devices has spearheaded the rapid development of molecular electronics, a field that aims to overcome the manufacturing limitation of silicon-based integrated circuits. In this context, several powerful experimental techniques have emerged that are dedicated to measuring the electronic properties of sub-nanometer scale single molecules. In this context, the all-electronic Single-molecule Break Junctions (SMBJ) have emerged as a leading technology [60], [75]. It enables the exploration of electronic properties at the scale of single molecules, such as a short strand DNA. The SMBJ technique offers a brand-new perspective of understanding the mechanics of quantum charge transport in single molecules and their potential applications in chemical structure analysis, biological marker detection, and nano-scale molecular electronics [60].

Current traces obtained from SMBJ experiment provide unique characteristics for identifying single molecules. Despite its groundbreaking potential, SMBJ faces significant challenges. The inherent noise in these traces is the major bottleneck, which stems from complex interactions between substrate, electrode, sample, environment, and measuring systems. If no molecule is

captured, the current trace of one SMBJ measurement would decay exponentially and contain most likely noises. Conversely, when molecules are present, the captured molecule would form an electrode–molecule–electrode structure, which measures the electronic characteristics of the molecule. Theoretically, these characteristics are not affected by small displacement changes. Therefore, one current trace would stay relatively constant, which forms a plateau [76], [77], [78], [79], [80], [81]. Among all raw current traces, these plateau segments contain the majority essential conductance features about the target molecule. These plateau conductance values are crucial for single molecules' identification with SMBJ [82], [83], [84].

Various methods are proposed to handle the extremely low signal-to-noise ratio of single-molecule measurements. Conventionally, for one molecule, thousands of iterations are required for a reliable analysis result. To segment the plateau, an empirical threshold could be directly applied to conductance traces [85]. In majority of studies, a 1D conductance histogram is constructed using thousands of current traces [56], [63], [65], [86], [87], [88] to capture the plateau conductance values. The conductance values of the target molecule could be represented with the location of peaks in a histogram (some of bins with highest counts). To accommodate the time (or distance between electrode) information along with conductance values, a 2D conductance histogram is another popular opinion [59], [62], [64], [66], [89], [90]. The resulting peak locations or plateau segments could provide a more accurate estimate of conductance value and distance measured. The high noise content in single-molecule measurements can be partially removed by averaging effect of 1D/2D histograms. With adequate large number of measurements, these methods could provide a reliable analysis result. In addition, various machine learning algorithms have been successfully applied to 1D/2D conductance histograms, providing a significant improvement of identification accuracy and efficiency [38], [73], [91], [92]. Some groups also tried to apply machine learning algorithms to extract the conductance plateau from an entire trace. For instance, using CNN-based weakly supervised learning with manual labeling the traces with plateau [73].

The drawbacks of constructing one conductance histogram directly from raw current traces are inevitable: including all plateau (signal) and decay (noise) segments, demand for large number of measurements, and lack of precise estimation of molecule conductance. Since the noise contents are occupying a large portion, conductance peaks become contaminated and distorted. Unless all the decay segments are removed, histograms cannot exhibit the real conductance features. In another word, it is highly challenging to quantify one of the few conductance values from a large-deviation non-prominent peak. To overcome these limitations, it becomes vital to develop a method for extracting the plateau segments from each raw conductance trace, without the need of manual labeling the plateau points. Although, there are limited analysis methods in the molecular electronics field [74]. Time-series analysis is a well-study topic for mathematics and statistics. Several works have developed algorithms to search linear segments iteratively until reach a local minimum or stopping criterion [93], [94], [95], [96]. After intensive investigation, we adopted the Optimal Piecewise-Linear Function

Approximation method [97]. We engineer this method to find an approximation that puts one conductance trace into several piecewise linear segments. Then, utilizing the slope of piecewise linear segments, we could easily isolate the plateau segments in the conductance traces. This method significantly enhances signal clarity and reduces the number of single-molecule measurements needed for reliable data.

1.5 Overview of Viral Testing Methods

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, and so on [21], [22]. 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 [98]. The ability to accurately identify mutations with single/few base-pair(s) mismatches holds particular significance in the context of viral testing, as evidenced during the recent COVID-19 pandemic.

COVID-19 (coronavirus disease 2019) is an infectious disease caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2). Since its first confirmed appearance at the end of 2019, COVID-19 has infected approximately 700 million people (10% of the world population) and resulted in the deaths of over 7 million people [99], [100]. Due to the virus's ongoing mutations, it is difficult to predict when the pandemic will come to an end. Each time the global weekly infection curve shows signs of decrease, a new COVID-19 variant with more transmissive or more virulence would take over and continue to break the infection records [99], [101]. Even if the pandemic ends, COVID-19 will likely become another recurrent disease that forever lives with the human species [102]. Abundant strategies have been developed to reduce transmission and stop the spreading of SARS-CoV-2: viral testing, contact tracing, and vaccination [103], [104].

Broadly speaking, viral testing can be categorized into three different scenarios: diagnostic testing, screening testing, and public health surveillance [105]. Diagnostic methods are employed for individuals exhibiting COVID-19 symptoms or those who have had recent close contact with symptomatic individuals. The most important criterion of diagnostic methods is accuracy, which needs high sensitivity and specificity at the same time. High sensitivity means few false-negative results, where individuals with COVID-19 are incorrectly identified as negative. High specificity means few false-positive results, where individuals without the virus are mistakenly identified as having COVID-19 [103], [105]. Screening methods, on the other hand, are used for individuals without COVID-19 symptoms and no known close contact with symptomatic individuals. Because the percentage of asymptomatic SARS-CoV-2 infections is very high (about 40%), screening methods are essential for identifying undetected infected individuals and reducing further transmission [106], [107]. Unlike diagnostic testing, screening methods prioritize repeatability and testing time over high accuracy [108]. Public health surveillance methods are

designed to gain information at a population level rather than an individual level, which means cost and accessibility are the most vital features [105], [109]. However, there is no existing viral testing method that fully meets the essential criteria for all three scenarios. Based on these scenarios, numerous viral testing methods have been developed, and more than hundreds of them are commercially available after clearing FDA-EUA (Emergency Use Authorization of the United States Food and Drug Administration) [110], [111], [112].

Viral testing can be broadly classified into two major categories: nucleic acid tests and protein tests (primarily antigen tests). Nucleic acid tests target some specific RNA (ribonucleic acid) genes in the virus. Among all nucleic acid tests, the RT-PCR (reverse transcriptase polymerase chain reaction) test is the most well-known method, which is highly accurate and widely available to the public. Several studies have shown that RT-PCR has nearly 100% sensitivity and specificity, making it the gold standard (above 99.99%) for viral testing [108], [112], [113]. However, the main drawback of RT-PCR remains its requirement for laboratory-based testing, resulting in turnaround times ranging from 1 to 3 days. Under high demand, it would take a week or even longer [103], [108]. While some newer nucleic acid tests offer shorter assay times (between 1 to 4 hours), their accuracy suffers (with sensitivity 87% - 94%) and becomes less reliable than the gold standard (above 99.99%) [112]. Therefore, RT-PCR is still the dominant method for the diagnostic test.

Protein tests in viral testing target either human antibodies or viral antigens. The human antibody tests are considered suboptimal for COVID-19 viral testing because of the possibility of corruption by residual antibodies from a previous infection or vaccination [103]. Among all protein tests, viral antigen tests have gradually become the most popular method, with the advantages of rapid turnaround time (15 minutes), low cost (about $10), and direct accessibility (at-home self-testing) [114]. While antigen-based methods sufficiently diagnose symptomatic patients, they face challenges in diagnosing asymptomatic patients. The sensitivity of these methods falls around 80% - 90%, leading to a large number of false-negative results. Consequently, confirmatory testing is often merited to avoid leaving COVID-19 carriers undetected and prevent further transmission of the virus [106], [107], [112]. However, no rapid viral testing method can compare to RT-PCR’s level of accuracy. Somewhat surprisingly, antigen-based methods remain the preferred choice for screening tests despite their limitations. Our method demonstrates that accurate classifiers can be built utilizing the electronic fingerprints of multiple COVID-19 variants and their single base-pair mismatches.

2 A Machine Learning Approach for DNA Sequence Identification

The all-electronic Single Molecule Break Junction (SMBJ) method is an emerging alternative to traditional polymerase chain reaction (PCR) techniques for genetic sequencing and identification. Existing work indicates that the current traces recorded from SMBJ experiments contain unique signatures to identify known sequences from a dataset. However, the current traces are typically extremely noisy due to the stochastic and complex interactions between the substrate, sample, environment, and measuring system, necessitating hundreds or thousands of experimentations to obtain reliable and accurate results.

In this chapter, we demonstrate that double-stranded DNA molecules can be classified extremely accurately using machine learning methods operating on experimental quantum transport data. Typical classification accuracies for molecules that are structurally different exceed 99.5%. Even in the case of DNA-RNA hybrids from E. coli with a single base pair mismatch, our methods are able to differentiate between them with an accuracy of over 96%. The overall accuracy over ten datasets of different molecules is 98.85%, with an average classification time of 27.49 μs (on a 2.2 GHz processor with 16 GB RAM) for each processed data input. However, we demonstrate that the accuracies of 'problematic' sequences such as S3 and S4 can be boosted to around 99.5% if a two-stage classifier (a multiclass model cascaded with a binary model) is employed instead of a single-stage classifier. Our analysis and simulation results demonstrate the potential of combining current traces and ML methods as a diagnostic tool for real-time detection and classification of genetic sequences.

2.1 Methodology

2.1.1 Data acquisition

All SMBJ experimental data are obtained directly from our collaborator, Josh Hihath's group. For a more detailed experiment setup, please refer to [16]. The SMBJ experiments are conducted at room temperature using Molecular Imaging Pico-STM. The STM probe is connected to a Digital Instruments Nanoscope IIIa controller which records the current traces. Subsequently, the current traces are converted to conductance traces.

For each experiment, a small volume of RNA:DNA hybrid molecule is first injected between two gold electrodes (the tip and substrate), meeting a desired concentration level. A bias voltage within the range of 50–300 mV is applied between the two gold electrodes. The STM probe, which is controlled by the LabView program, is then moved towards the substrate at a rate of approximately 80 nm/s until the current saturates the preamplifier (~100 nA). The tip is then retracted at the same rate until the current reaches the lower limit of the preamplifier (~10 pA). The whole process is repeated until the tip reaches the edge of the substrate. Among the thousands of collected traces, a significant number shows a predominantly exponentially decaying form, which implies that no molecules were captured between the tip and the substrate

(invalid experiment). A small number of current traces show steps and flat regions (in general, substantial deviations from an exponentially decaying characteristic), which is indicative of a successful molecular binding between the tip and the substrate.

The SMBJ approach consists of a molecule binding to a conducting substrate on one end and a scanning microscope tip on the other end. The current flow between the tip and substrate is measured as a function of time. Figure 2-1 (a) shows an idealized schematic of the SMBJ experimental setup. Representative current traces without and with DNA molecular binding between the gold electrode and the substrate are shown in Figure 2-1 (b) and (c) respectively. In the absence of any binding, the current trace exhibits a predominantly exponential decay as the electrode is moved away from the substrate. Deviations from this behavior as shown by increases/dips in Figure 2-1 (c) are generally indicative of a successful molecular binding and a 'valid' experiment. Our hypothesis is that unique signatures of the DNA molecule exist in the conductance (ratio of current to applied voltage) traces, specifically, conductance histograms (see Figure 2-6 and Figure 2-7), and automated classification of DNA strands should be possible using statistical and/or machine learning (ML) based approaches.

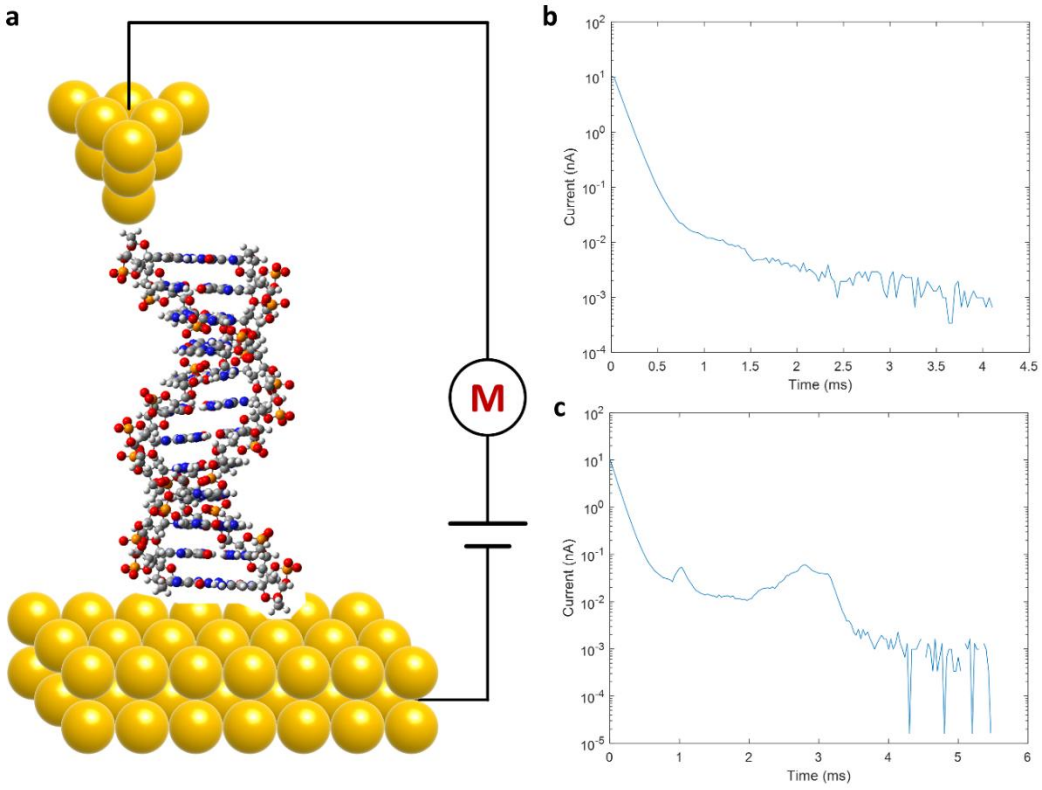


Figure 2-1. Raw experimental input data. (a) An idealized schematic of the single-molecule break junction (SMBJ) approach for conductance measurement. The measurement setup includes a bias voltage and a preamplifier [115]. (b) Sample current trace with no DNA binding between the gold electrode and the substrate. (c) Sample current trace with DNA binding between the gold electrode and the substrate.

We use ten datasets of experimentally obtained current traces to validate our ML approach. Each dataset contains a mix of valid (with molecular binding) and invalid (no molecular binding) traces. In the next section, we explain how the invalid traces were automatically pre-filtered prior to classifier training. Table 2-2 provides the details of the ten datasets. Some salient features of the datasets are as follows. First, although S2, S6, S7, S8, and S9 are of the same strand, tests were conducted using three different bias voltages. Second, S4 and S5 are mismatched versions of S3, which corresponds to mRNA from E. coli: O157:H7 with its fully matched DNA duplex and is known to produce both Shiga toxins (Stx) 1 and 2. S4 and S5 have the same DNA complement as S3. The mRNA from S4 corresponds to E. coli O175:H28. It has a single mismatch at base 14 of S3 (A is substituted by G) and is known to lead to Stx 2. The mRNA from S5 which corresponds to E. coli E1a, has a single mismatch at base 8 of S3 (C is substituted by T) and is nontoxic as it does not produce either Stx 1 or 2.

2.1.2 Data pre-processing

A series of pre-processing steps were used to convert the experimentally obtained SMBJ current traces to conductance histograms for training our ML models. Since the recorded current values are the outputs of a preamplifier, which has a noise floor of 10 *p*A and an upper limit of 100 *n*A, all current traces were first limited to the threshold range [10 *p*A – 100 *n*A].

Due to uncertainties in the data acquisition process, sometimes an experiment is conducted even when no molecule is attached to the electrode. Such experiments are deemed to be invalid. Invalid current traces exhibit a predominantly exponential decay over time. Since our datasets had a mix of valid and invalid traces for each molecule, we designed the second pre-processing step. We fit an exponential regression model, $y = a \cdot e^{-b \cdot x},\ b > 0$, to each clipped current trace and used the R^2 statistic from the fit to accept/reject the trace. 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. 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. Starting with about 1400–7000 traces for each of our 10 data classes, we found that approximately 50–90% of the data were rejected at this step for β = 0.95, 70-96% of the current traces were rejected for β = 0.90, and 78-98% of the current traces were rejected for β = 0.87 (see Table 2-1). Given the scarcity of experimental data, a prudent choice of β is necessary to retain as much data as possible, while rejecting the most obviously invalid current traces.

In the third step, all accepted current traces were low pass filtered (LPF). Two different LPFs were used, one for data classes S1–S9 with a cutoff frequency of 9 kHz and another for data class S10 with a cutoff frequency of 3 kHz. Since the sampling rate during the data acquisition phase for S10 was 10 kHz, compared to 30 kHz for S1–S9, all S10 current traces were linearly interpolated by a factor of 3 after low pass filtration.

Table 2-1. Data availability after R^2 test*

	$R^2 \leq \beta = 1$	$R^2 \leq \beta = 0.95$		$R^2 \leq \beta = 0.9$		$R^2 \leq \beta = 0.87$	
S1	5254	528	10.05%	176	3.35%	103	1.96%
S2	5008	1212	24.20%	729	14.56%	571	11.40%
S3	7008	1997	28.50%	872	12.44%	532	7.59%
S4	6187	2347	37.93%	1254	20.27%	874	14.13%
S5	5420	1804	33.28%	926	17.08%	604	11.14%
S6	4998	2280	45.62%	1497	29.95%	1142	22.85%
S7	1369	538	39.30%	299	21.84%	206	15.05%
S8	4171	2034	48.77%	1097	26.30%	831	19.92%
S9	4719	1697	35.96%	856	18.14%	601	12.74%
S10	5037	1578	31.33%	938	18.62%	670	13.30%

* Columns 2, 3, 5, and 7 show the number of data samples in each dataset at different values of the R^2 test threshold parameter, β. Columns 4, 6, and 8 show the percentage of samples remaining.

Finally, all current traces were converted to conductance traces using eq. below:

$$Conductance = \frac{current}{V_{bias}} \tag{2-1}$$

where current is in units of 10 nA and V_{bias} is the bias voltage in volts. The unit of conductance is G_0, the *conductance quantum*, and is defined as follows: $G_0 = \frac{2e^2}{h} = 7.748 \cdot 10^{-5}\ S$, where e is the elementary charge and h is Planck's constant. For our datasets, the conductance values turn out to be in the range $[10^{-6.5} - 10^{-0.5}]\ G_0$.

Figure 2-2 summarizes the sequence of pre-processing steps. Training vectors for our ML models are probability histograms computed from sample conductance traces. The efficacy of the pre-processing routine in smoothing out the histograms is demonstrated in Figure 2-3 which compares the conductance histograms for data class S1 computed from raw current traces vs. processed traces.

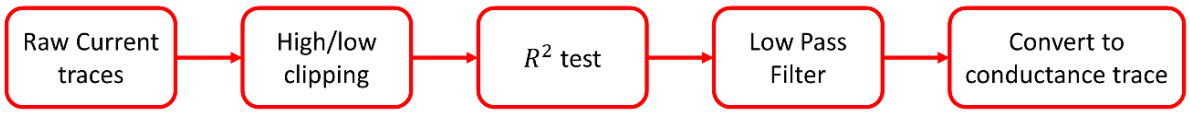


Figure 2-2. Sequence of pre-processing steps for converting raw current traces to conductance traces.

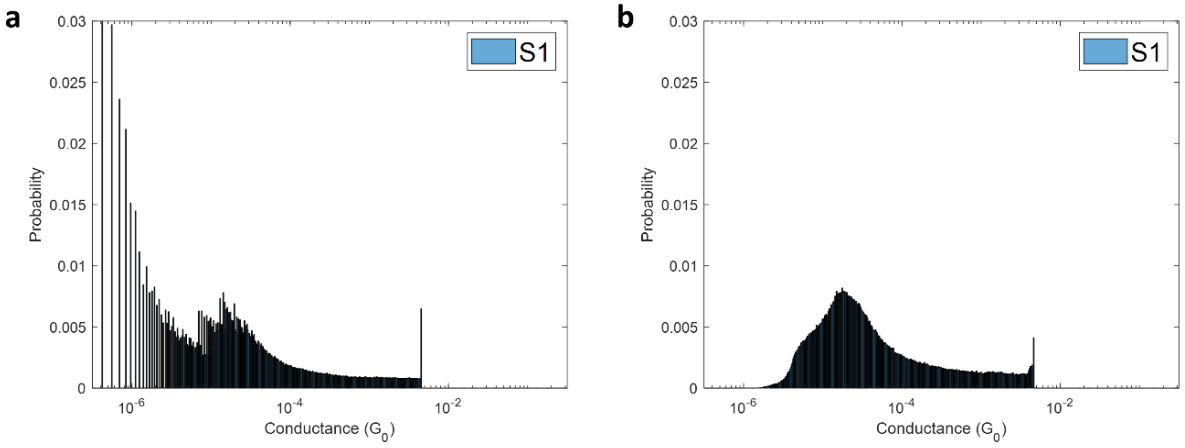


Figure 2-3. Representative large sample conductance histograms for data class S1 using threshold parameter β = 1. (a) Derived from raw current traces. (b) Derived from processed current traces.

We used the Python implementation of the XGboost package available at [116]. Two critical hyperparameters within XGboost are the number of trees/estimators, N_{est}, and the depth of each tree/estimator, D_{est}. Optimal values of these parameters were determined based on an exhaustive grid search on values indicated below:

D_est = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 18, 20}
N_est = {10, 50, 100, 150, 200, 250, 300, 350, 400, 500, 600, 700}

Based on validation accuracies, we determined that reasonable values are (N_{est} = 200, D_{est} = 2) or (N_{est} = 150, D_{est} = 3). For all simulations, we chose the first parameter combination to reduce the computational time. All other parameters were left at their default values. In particular, the default value of the learning rate or shrinkage parameter is 0.3.

2.1.3 Data visualization

The *t*-stochastic neighbor embedding (*t*-SNE) method [117] is a powerful *local structure-preserving* mapping method to visualize high-dimensional data in two dimensions. A key parameter in *t*-SNE is 'perplexity', which controls the number of neighbors used by the algorithm to determine the local structure of a point. It is important to observe that not much should be made of the exact distance by which two distant points may be displayed on the low dimensional map generated by *t*-SNE since the algorithm is agnostic to 'global distance information' in the native feature space. An interesting discussion regarding the interpretability of *t*-SNE maps can be found in [118]. A recent application of the *t*-SNE algorithm to single-cell transcriptomics can be found in [119]. Figure 2-4 shows the *t*-SNE visualizations of our ten 600-dimensional conductance probability histogram datasets (1000 per class) for β = 1, 0.95, and 0.87, based on a Barnes-Hut approximation with perplexity value = 30 and an initial 100-dimensional principal components analysis (PCA) projection. Overall, there appears to be a

separation between the points in the different datasets. However, three important observations are in order.

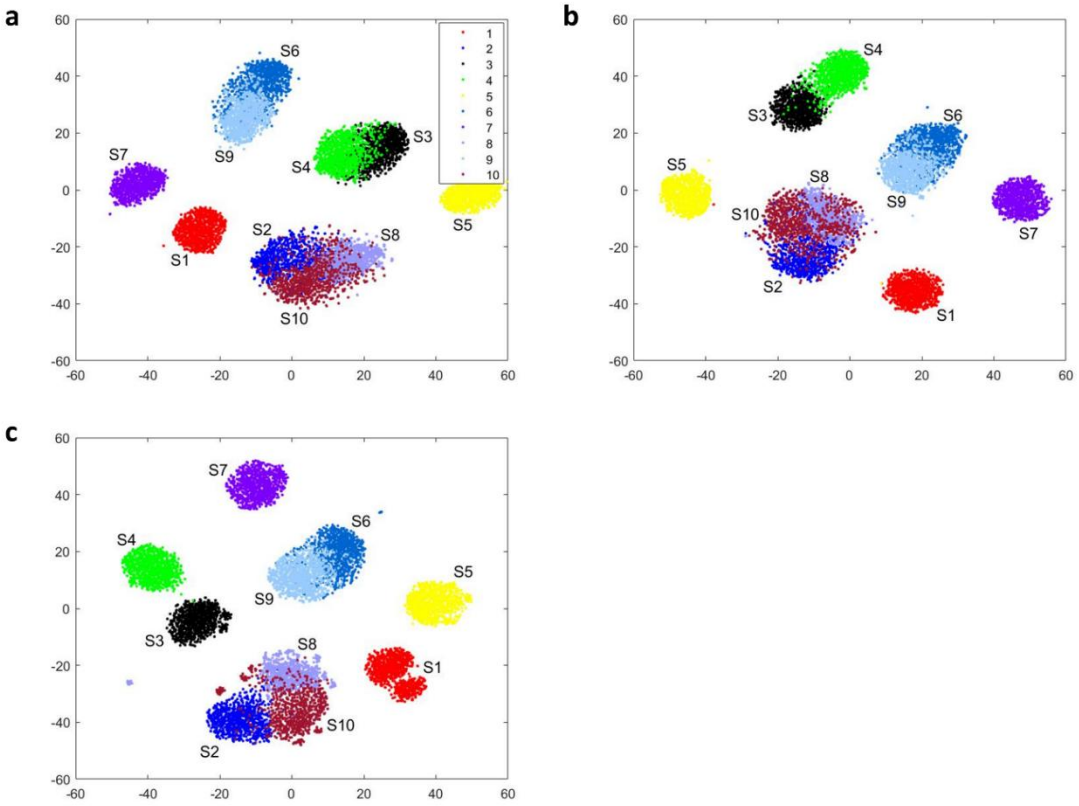


Figure 2-4. *t*-SNE visualization of our datasets. (a) $R^2 \leq \beta = 1$, (b) $R^2 \leq \beta = 0.95$, and (c) $R^2 \leq \beta = 0.87$. We have adopted the same 'bluish' color scheme for S2, S6, S7, S8, and S9 since they correspond to the same molecule, though with three different biases.

First, the separation between S3 and S4, which differ by a single mismatch, appears to improve as β decreases. As far as these two classes are concerned, we should therefore expect a classifier to benefit from an aggressive R^2 test-based pre-filtering of the data.

Second, the choice of bias voltage matters. This is evident from an examination of the clusters depicted by (S2,S8), (S6,S9), and S7. Although these datasets correspond to the same strand, we see from Table 2-2 that the bias voltage is 10 *m*V for (S2,S8), 100 *m*V for (S6,S9), and 200 *m*V for S7. Experiments conducted with different bias voltages on the same molecule can induce very different conductance properties, which is evident in the histograms in Figure 2-6 as well as in the natural separation we observe between (S2,S8), (S6,S9) and S7 in Figure 2-4. This lends some empirical evidence that a target class labeling scheme based on a (strand, bias voltage) tuple may be a more prudent choice than a purely strand-based labeling scheme which may result in a diffuse cluster of identically labeled points occupying a wide swathe in the feature space, which can be detrimental to an ML-based classification approach.

Third, we observe that the S10 cloud map 'straddles' the (S2,S8) cloud considerably and no amount of R^2 filtering appears to ameliorate this issue, unlike S3 and S4. In fact, without the aid of class labels, it would be natural to expect one cluster formed by S2, S8, and S10. It is interesting to note that our first and third observations regarding (S3,S4) and (S2,S8,S10) are also validated by classical multidimensional scaling [120] projection plots (see Figure 2-5).

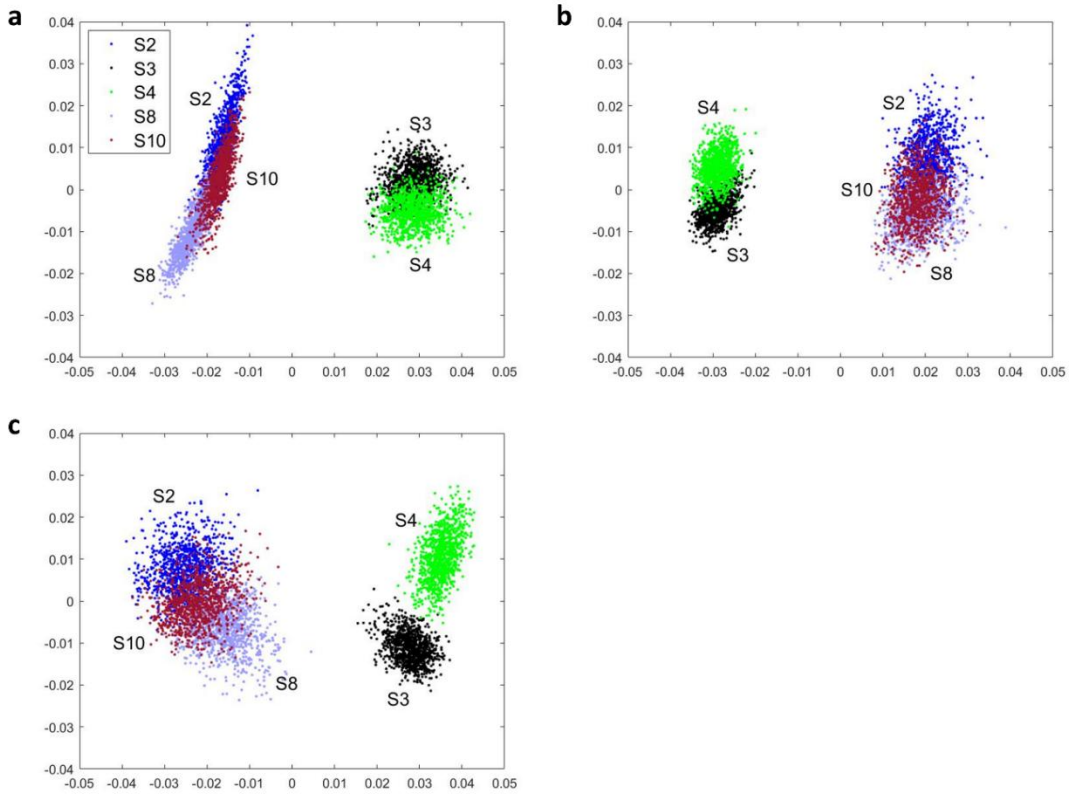


Figure 2-5. Classical multidimensional scaling (MDS) visualizations. MDS visualizations of S2, S3, S4, S8, and S10 datasets for (a) $R^2 \leq \beta = 1$, (b) $R^2 \leq \beta = 0.95$, and (c) $R^2 \leq \beta = 0.87$.

2.2 Results and Discussion

2.2.1 Overview of classification approach

Raw current traces from SMBJ experiments [16] (see Figure 2-1) are first converted to conductance traces through a series of pre-processing steps, which are discussed in Section 2.1.2. Histograms are then derived from the conductance traces. Figure 2-1 shows the empirical *large sample* conductance histograms for our ten datasets (see Table 2-2 for details) using an R^2 test threshold of $\beta = 0.95$. We refer to these histograms as large sample histograms since these are based on thousands of conductance traces for each dataset (see Table 2-1 for number of traces). From a practical perspective, however, strand-level identification needs to be made from reasonably accurate histograms constructed from the fewest number of conductance traces possible. This is necessary to avoid imposing a costly experimental burden, which can

potentially detract from the adoption of current-based methods for applications that demand almost real-time genetic sequence determination.

Table 2-2. Details of the ten datasets used for baseline classifier.

Label	Sequence	Bias	Note	Solvent / Buffer
S1	Octane dithiol	0.30 V	Not a DNA/RNA	mesitylene
S2	5'-CCC GGG CCC GGG-3' 3'-GGG CCC GGG CCC-5'	0.01 V		100mMP+100uL +30uL
S3	5'-CGA CCC CTC UUG AAC-3' 3'-GCT GGG GAG AAC TTG-5'	0.05 V	E. coli O157:H7	10uL+75uM+60 0MC_50BC_20 RR
S4	5'-CGA CCC CTC UUG A**G**C-3' 3'-GCT GGG GAG AAC TTG-5'	0.05 V	E. coli O175:H28 One mismatch from S3	30ul+7.5uM+Rg
S5	5'-CGA CCC C**C**C UUG AAC-3' 3'-GCT GGG GAG AAC TTG-5'	0.30 V	E. coli ED1a One mismatch from S3	75uM+10uL
S6	5'-CCC GGG CCC GGG-3' 3'-GGG CCC GGG CCC-5'	0.10 V	Same as S2	100mMP+100uL +30uL
S7		0.20 V		100mMP+100uL +30uL
S8		0.01 V		100mMP+100uL +20uL
S9		0.10 V		100mMP+100uL +50uL
S10	5'-GGG TTT GGG-3'	0.01 V	G-quadruplex secondary structures	100mMP+100uL +30uL

Ideally, we would like to conduct a successful (with molecular binding) experiment once and be able to predict the strand from the resulting conductance values. However, given the substantial noise and uncertainty inherently associated with the current state of SMBJ experimentation, a compromise is necessary. We train and evaluate our classifiers based on *H*-sample conductance histograms - histograms constructed from randomly sampled *H* 'valid' conductance traces (whether a trace is deemed valid or not is determined by the R^2 test threshold β) - and study the impact of the parameter *H* on classifier accuracy. Henceforth, we will refer to histograms constructed from reasonably small values of *H*, *e.g.* $H \leq 20$, as *small sample* histograms. Additionally, we define any histogram based on $H = 30$ as a *baseline* histogram. Figure 2-7 shows a representative large sample, baseline, and small sample ($H = 10$) histograms

for datasets S1 and S8. We observe that while the conductance distribution of S8 changes significantly at $H = 30$ (note the multiple peaks in the conductance range $[10^{-4}, 10^{-2}]\ G_0$), the distribution of the baseline histogram for S1 appears to be relatively consistent with its large sample counterpart. The significant change in the conductance histogram for S8 at $H = 30$ probably alludes to unique contact configurations between DNA and electrodes. Nevertheless, expecting consistently perfect experimental conditions is a luxury in practice, and classifier models need to be relatively robust to such uncertainties.

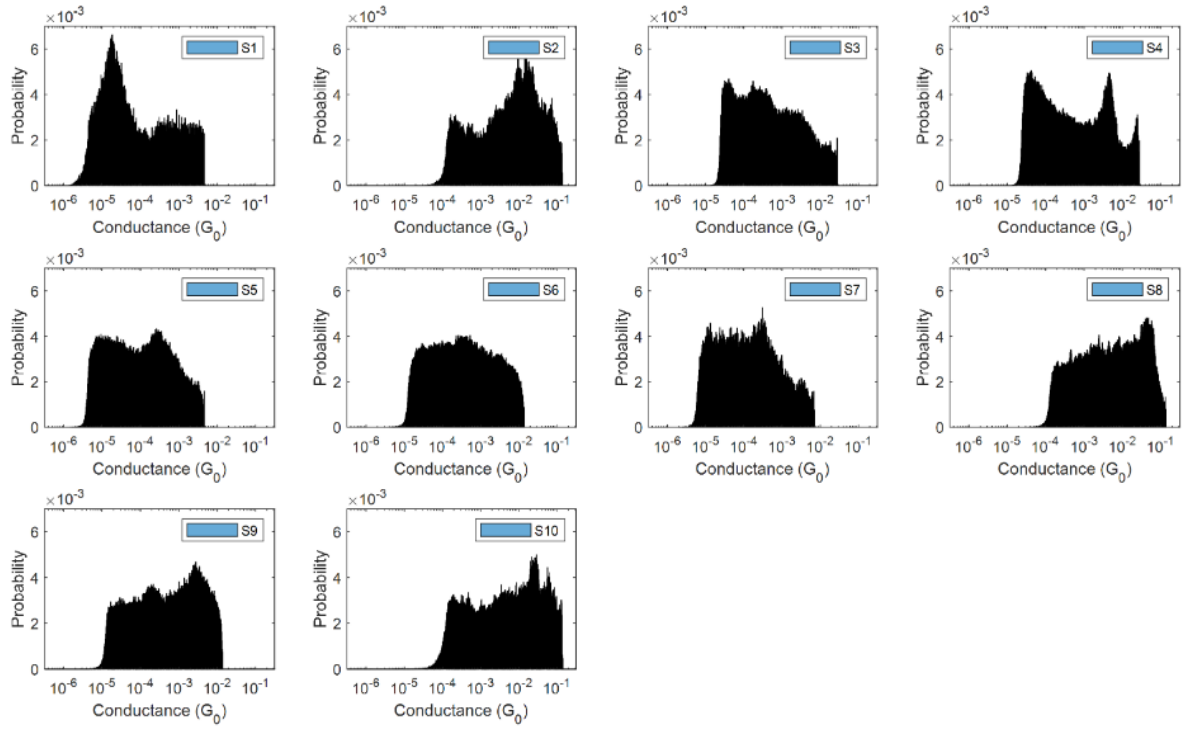


Figure 2-6. Large sample histograms of all data classes. The R^2 test is: 'accept current trace if $R^2 \leq \beta = 0.95$'. After the current traces are R^2 filtered, one histogram is constructed per class, using all available current traces, which are converted to conductance traces after low pass filtration (see Figure 2-2 and Figure 2-3 for details).

Starting with the ten datasets, we trained classifier models using two different target class labeling schemes. In the first scheme, unique DNA strands were assigned different class labels, irrespective of the voltage bias used for current measurement during SMBJ experiments. Since data types S2, S6, S7, S8, and S9 are of the same DNA strand, we end up with six target classes, [S1, {S2, S6, S7, S8, S9}, S3, S4, S5, S10], with this labeling scheme (called TLS-1). In the second scheme (called TLS-2), the datasets are assigned unique class labels based on the (strand, voltage bias) tuple. Although S2, S6, S7, S8, and S9 pertain to the same DNA strand, experimentations on S2 and S8 were conducted with a 10 *m*V bias, in contrast to 100 *m*V for S6 and S9 and 200 *m*V for S7 (see Table 2-2). Based on the conductance histograms shown in Figure 2-6, *prima facie* there appear to be enough differences induced by the choice of the bias

voltage, to warrant consideration of a labeling scheme based on (strand type, bias voltage). With this scheme, we have eight target classes, [S1, {S2, S8}, S3, S4, S5, {S6, S9}, {S7}, S10].

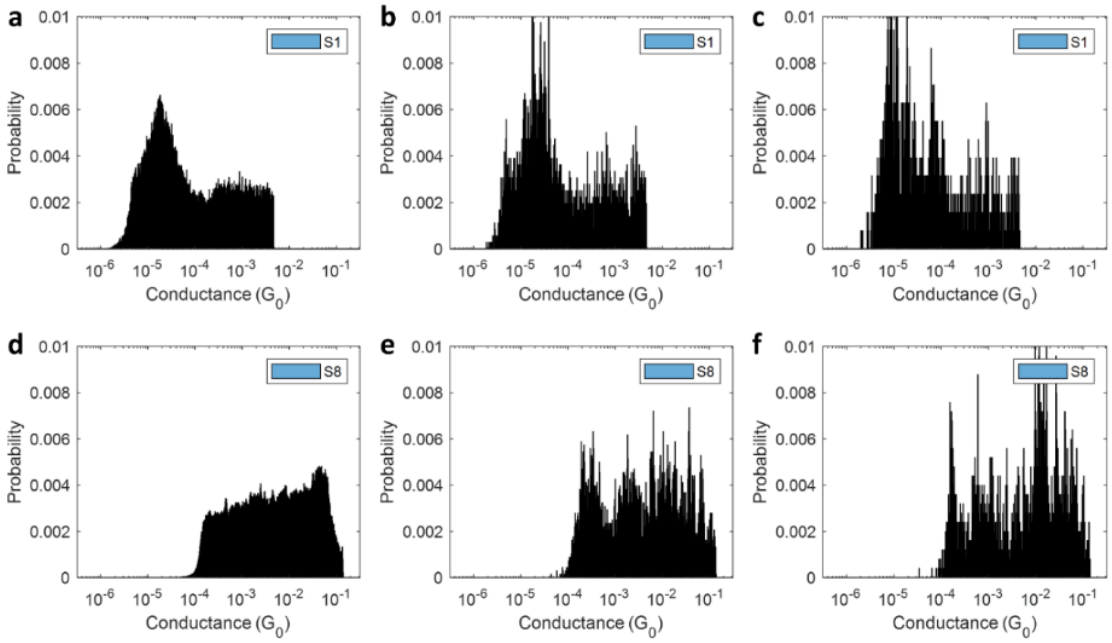


Figure 2-7. Comparison of large sample, baseline, and small sample histograms. (a), (d) Large sample, (b), (e) baseline (H = 30) and (c), (f) small sample (H = 10) conductance probability histograms for datasets S1 (a-c) and S8 (d-f).

For both labeling schemes, we randomly partitioned our conductance traces, 70% for training and 30% for testing. To reduce any potential bias due to the initial split, each simulation was repeated 100 times with a random train/test split. For each simulation, H traces were sampled randomly from each class to generate a conductance histogram. This process was repeated for a total of N_{hist} = 700 times per class, where N_{hist} is the number of histograms. The total number of training samples is, therefore, 4200 for TLS-1 (six classes) and 5600 for TLS-2 (eight classes). A similar procedure was adopted for evaluating classifier performance, except we chose N_{hist} = 300 per class. The total number of testing samples is therefore 1800 for TLS-1 and 2400 for TLS- 2. All accuracy metrics reported in the following sections are averaged over 100 random test splits.

The histograms for each target class can be constructed either from the set of conductance vs. time traces, $g(t)$, or the set of corresponding conductance magnitude trace, $|G(\omega)|$, where $|G(\omega)|$ is the magnitude of the Fourier Transform of $g(t)$. Henceforth, we will refer to $|G(\omega)|$ simply as the conductance spectrum. We evaluate the performance of four different classifier models:

- Approach 1: (see Figure 2-8) Extreme gradient boosting (XGboost) on 600-bin histograms constructed from H conductance traces sampled randomly from the set $g(t)$.

- Approach 2: (see Figure 2-9) XGboost on 600-bin histograms constructed from H conductance traces sampled randomly from the set $|G(\omega)|$.
- Approach 3: (see Figure 2-10) Multilayer perceptron (MLP with two hidden layers, $600-64-8(6)-8(6)$ with ReLu activation; numbers within parentheses refer to TLS-1) on 600-bin histograms constructed from H conductance traces sampled randomly from the set $g(t)$. From our experience, shallower networks tended to perform better than deeper networks, possibly due to the inherently noisy nature of the conductance traces.
- Approach 4: (see Figure 2-11) A distance-based classification scheme on 600-bin histograms constructed from H conductance traces sampled randomly from the set $g(t)$. For each target class, one '*large sample template histogram*' is constructed using all training data for that class. A test histogram (constructed from H samples), say t, is assigned to class i if the Euclidean distance between t and H_i is the smallest, where H_i is the large sample template histogram of the i^{th} class.

Since the performances of other approaches are generally at par with or slightly worse than approach 1, we will concentrate on the performance of the boosted method in the next section. Broadly speaking, XGboost [121] is a fast and scalable implementation of a gradient-boosted decision tree framework [122]. Gradient boosting is an ensemble learning method wherein weak base learners (usually decision trees) are added sequentially, one at each iteration, to minimize a suitably defined loss function evaluated on the previous learner. Within XGboost, the loss function is cross entropy for multiclass classification problems. For details on gradient boosting in general, we refer the reader to Hastie's article [120]. We used the Python implementation of the XGboost package [116]. Two critical hyperparameters within an XGboost framework are the number of trees/estimators, N_{est}, and the depth of each tree/estimator, D_{est}. Based on extensive hyperparameter optimization, we chose $N_{est} = 200$ and $D_{est} = 2$.

Let N_{bins} denote the number of bins for a conductance probability histogram. We define a *baseline classifier* as one which is characterized by the following parameters: (i) $N_{bins} = 600$ (*i.e.*, each training sample is a 600-dimensional probability vector), (ii) $R^2 \leq \beta = 0.95$ in the pre-processing sequence, and (iii) $H = 30$.

In the subsequent Section 2.2.2-2.2.5, we first discuss the performances of the baseline classifiers for both target class labeling schemes, followed by an analysis of their sensitivities to the choice of N_{bins}, β, and H. Irrespective of the choice of N_{bins}, β, and H, we have consistently used $N_{hist} = 700$ per class for training and $N_{hist} = 300$ per class for testing. All classifier accuracy results reported are averaged over 100 evaluations. Finally, we discuss a secondary binary classifier for S3 and S4 in Section 2.2.6. We refer the reader to Section 2.1.3 for *t*-SNE and multidimensional scaling maps of our data. These low-dimensional visualizations provide broad insights into the structure of the data and are useful in deriving insights into classifier performance.

2.2.2 Performance of baseline classifiers

Figure 2-8 shows the confusion matrices of the baseline classifiers for both target class labeling schemes. We recall that datasets S2, S6, S7, S8, and S9 pertain to the same DNA strand, though with three different bias voltages. First, we observe that classifier accuracies for these five datasets are not affected by the choice of the labeling scheme. Highly accurate strand-level identification is possible whether the data is labeled based purely on strand type or a finer (strand type, bias voltage) tuple. Any ML framework should be able to utilize data from instrumentation that uses less sensitive current amplifiers. Lower sensitivity current amplifiers require a larger bias to record current. As the bias increases, molecular conformation may change in manners that are not yet fully understood or characterized. Consequently, the ability to detect a strand independent of the bias applied is invaluable. Second, we observe that classes S3 and S4 can be differentiated with an accuracy exceeding 96.5%, despite the fact that S4 is just a single mismatch mutant of S3. Although TLS-2 seems to enjoy a slight edge in performance as far as S3 and S4 are concerned, the difference is not appreciable to claim a definite advantage for any particular labeling scheme. Third, we note that class S5 poses no issues for the baseline classifiers and is always identified perfectly. Fourth, whenever (S2, S8) is misclassified, it is predicted to be S10, and vice versa. This is consistent with the low-dimensional visualization shown in Figure 2-4 and Figure 2-5. Finally, we observe that, barring S3 and S4, the XGboost baseline classifiers work phenomenally well with individual class accuracies exceeding 99.5% for $H = 30$. For S3 and S4, the accuracies are within 96.2-96.7%. Our investigations indicate that instead of a single classifier, a tandem classifier architecture, consisting of a primary classifier involving all classes and a secondary binary classifier operating only on S3 and S4 (see Table 2-3), can significantly boost the prediction accuracies of classes S3 and S4 to approximately 99.5%. Additional discussions are available in Section 2.2.6.

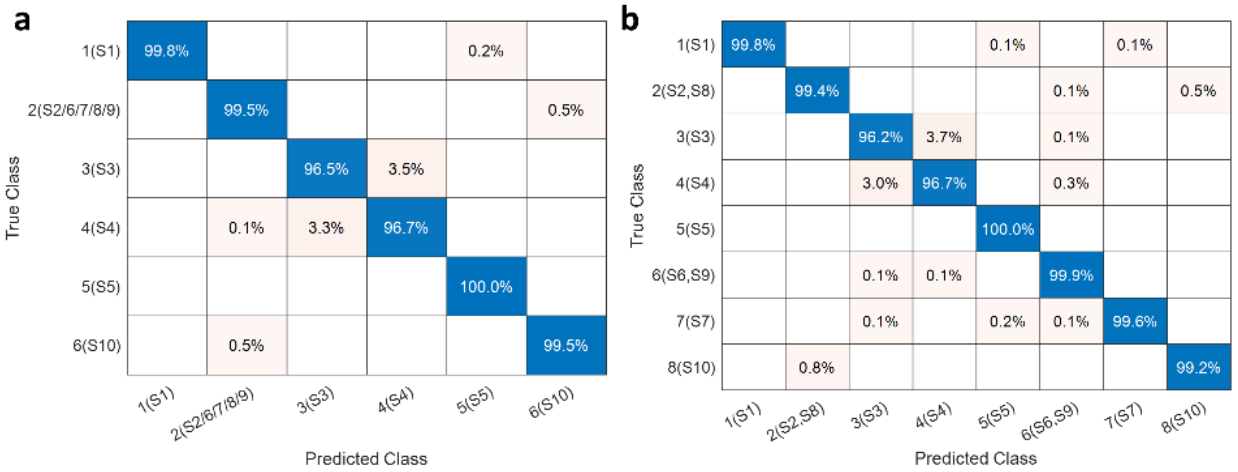


Figure 2-8. Confusion matrices for baseline classifiers. (a) Confusion matrices corresponding to target labeling scheme TLS-1 with 6 classes. (b) Confusion matrices corresponding to target labeling scheme TLS-2 with 8 classes.

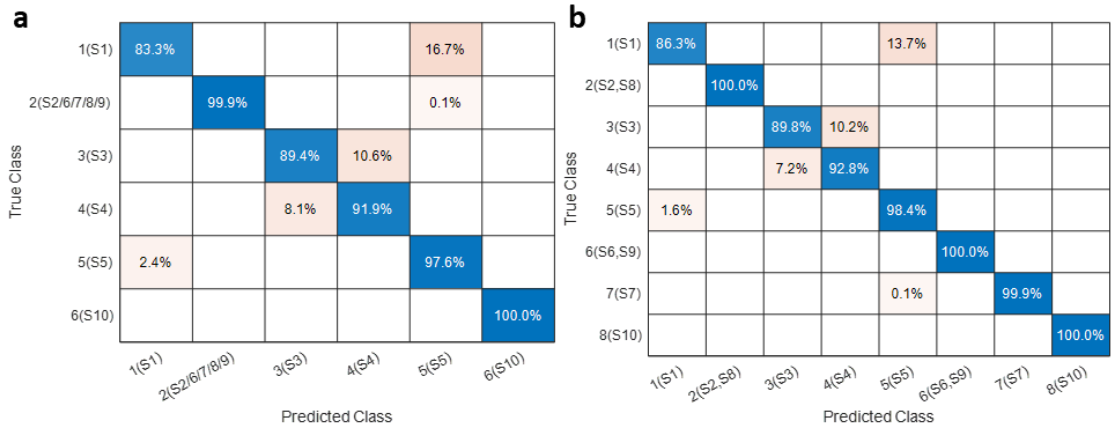


Figure 2-9. Confusion matrices for XGBoost classifiers with magnitude of the Fourier Transform. (a) Confusion matrices corresponding to target labeling scheme TLS-1 with 6 classes. (b) Confusion matrices corresponding to target labeling scheme TLS-2 with 8 classes.

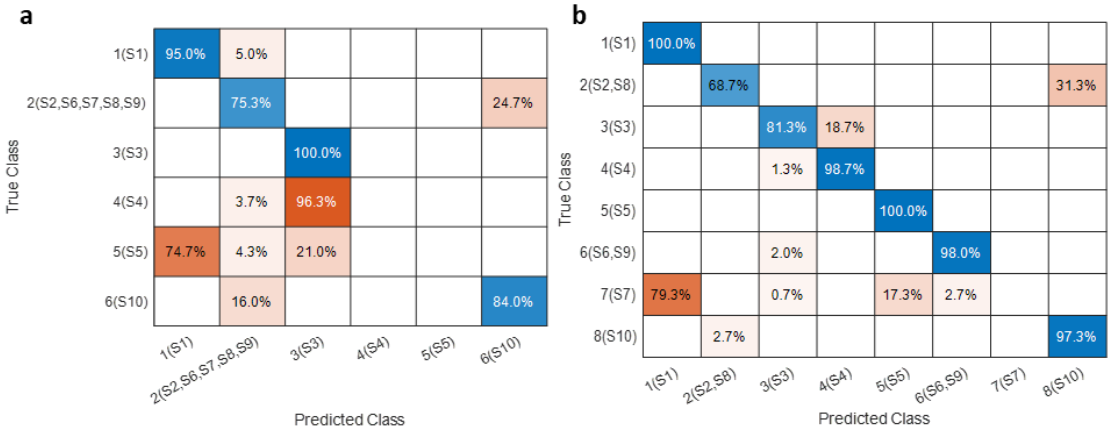


Figure 2-10. Confusion matrices for Multilayer perceptron classifiers. (a) Confusion matrices corresponding to target labeling scheme TLS-1 with 6 classes. (b) Confusion matrices corresponding to target labeling scheme TLS-2 with 8 classes.

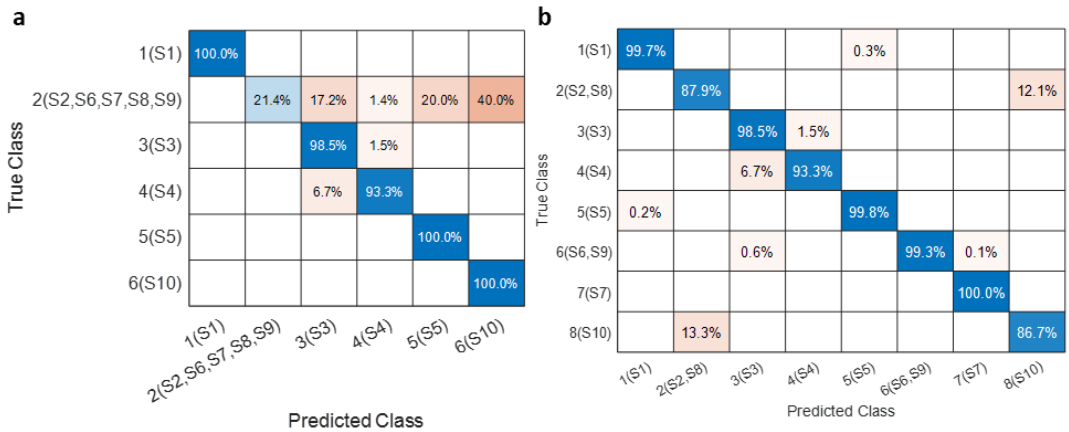


Figure 2-11. Confusion matrices for distance-based classifiers. (a) Confusion matrices corresponding to target labeling scheme TLS-1 with 6 classes. (b) Confusion matrices corresponding to target labeling scheme TLS-2 with 8 classes.

2.2.3 R^2 test threshold parameter

First, we focus on the performance analysis of baseline classifiers with respect to the R^2 test threshold parameter, β. Figure 2-12 (a) and (b) show the accuracy of the baseline classifiers with respect to (w.r.t) the R^2 test threshold parameter, β, corresponding to the two different target labeling schemes, maintaining N_{bins} = 600 and H = 30 (for detailed confusion matrices, see Appendix Figure A-1). Intuitively, we should expect a higher number of invalid (no molecule) or somewhat invalid (traces that are substantially similar to an exponential decay) traces to be rejected for lower values of β. Provided an adequate number of traces are retained after this step, this should yield a clean pool of valid traces to sample from, which should translate to improved or similar classifier accuracies. However, we observe that the accuracy for S1 drops significantly for lower values β with both labeling schemes, as does the accuracy for S7 for TLS-2. This can be explained by the fact that only 103 and 207 current traces remain for S1 and S7 respectively, when β = 0.87. Even though N_{hist} = 700 for all classes during training, the shallow pool of traces to sample from means that a substantial number of the training histograms for these two classes are reasonably similar and do not contribute to classifier learning. On the other hand, datasets S3 and S4 show an almost 8% improvement in accuracy as β is lowered from 1.0 to 0.87. This result can be further understood by observing the *t*-SNE visualization (see Figure 2-4), which shows a mostly clean separation between the S3 and S4 clouds at β = 0.87. At the lower end of β, 532 and 874 current traces remain for S3 and S4 respectively, enough for meaningful repeated sampling. We postulate that for reasonably high values of β, small/medium sample training histograms of these two classes are somewhat contaminated by invalid or substantially invalid traces, which masks out the true differences we expect between the two classes. A lower value of β which aggressively filters out raw traces, decreases the odds of histogram contamination, and allows the classifier to discover subtle differences, leading to a substantial enhancement in accuracy. Therefore, we conclude that a '*one β fits all*' policy need not be the most prudent choice for all classes of data.

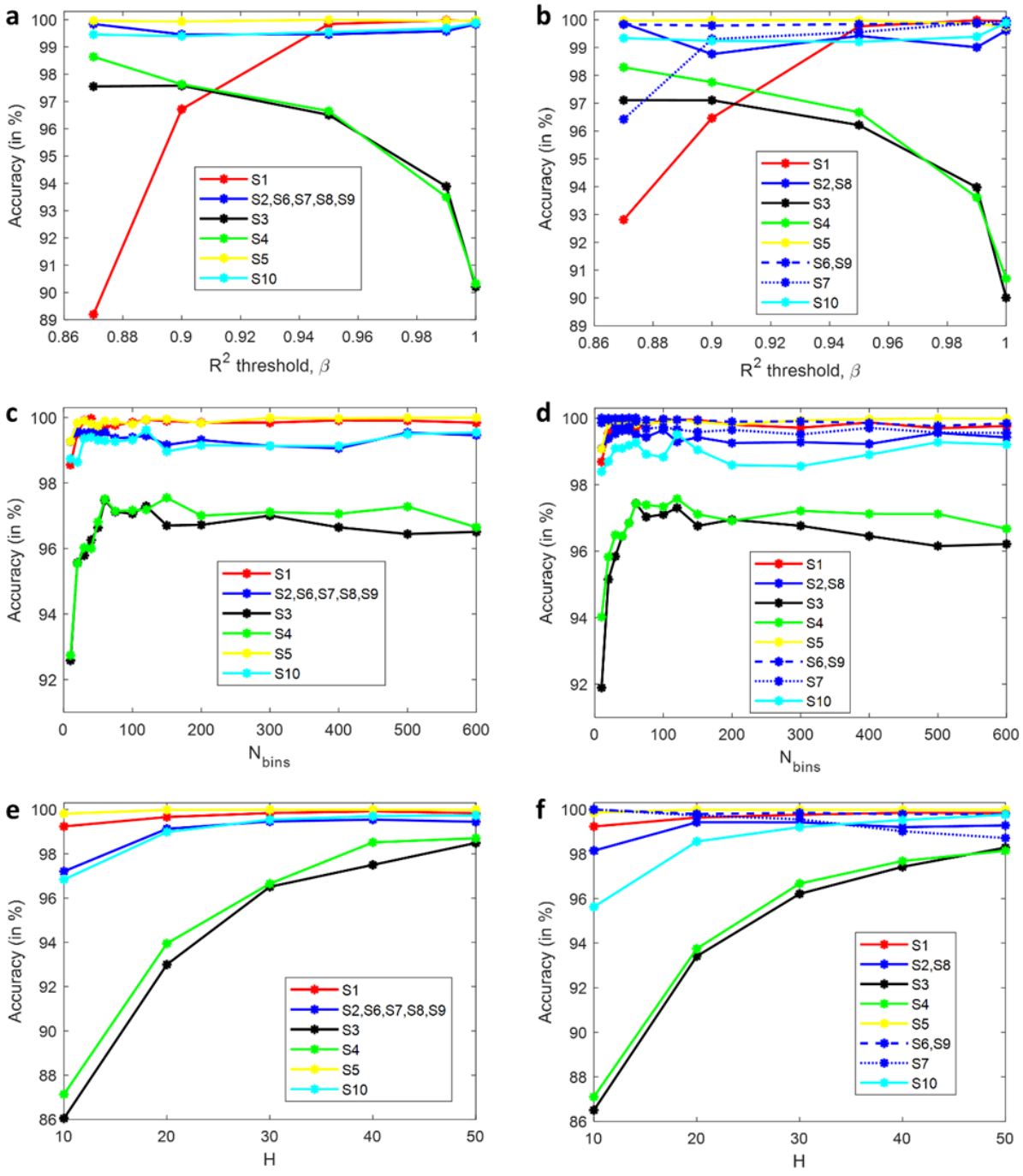


Figure 2-12. Performance analysis of baseline classifiers with respect to (w.r.t) β, N_{bins}, and H. (a), (b) Accuracy of baseline classifiers w.r.t the R^2 test threshold parameter, β, corresponding to target labeling scheme TLS-1 with 6 classes (a) and TLS-2 with 8 classes (b). We chose the same color but different line types to distinguish between datasets [S2,S8], [S6,S9], and [S7], which are of the same strand but use different bias voltages.
(c), (d) Accuracy of baseline classifiers w.r.t the number of histogram bins, N_{bins}, corresponding to target labeling scheme TLS-1 with 6 classes (c) and TLS-2 with 8 classes (d). Similar color

schemes and line types as in (a, b) have been used to distinguish between datasets [S2,S8], [S6,S9], and [S7].
(e), (f) Accuracy of baseline classifiers w.r.t the number of traces used to compute a conductance histogram, *H*, corresponding to target labeling scheme TLS-1 with 6 classes (e) and TLS-2 with 8 classes (f). We chose the same color but different line types to distinguish between datasets [S2,S8], [S6,S9], and S7, which are of the same strand but use different bias voltages.

2.2.4 Number of histogram bins

We then focus on performance analysis of baseline classifiers with respect to the number of histogram bins, N_{bins}. Figure 2-12 (c) and (d) show the accuracy of the baseline classifiers w.r.t the number of histogram bins, N_{bins}, corresponding to the two different target labeling schemes, maintaining $\beta = 0.95$ and $H = 30$ (for detailed confusion matrices, see Appendix Figure A-2). Typically, histograms with small bin widths are preferred because they closely resemble the underlying probability distribution. This is affirmed in Figure 2-12 (c) and (d), where we observe that the accuracies are generally highest when the number of bins is large (600), except for classes 3 and 4. For these two classes, we observe that a coarse binning strategy with just 60 bins has better accuracy for both labeling schemes, which is evidence that '*finer is not always better*' when it comes to choosing the granularity of the conductance histograms. While we do not have a definite explanation for why a smaller number of bins is better for S3 and S4, we conjecture that a fine binning strategy, especially when coupled with smallish values of *H*, generates 'pseudo-features' that are detrimental to learning by overfitting due to *curse of dimensionality*. We conjecture that a coarse binning strategy with its inherent noise averaging properties appears to avoid this issue and may be a better choice for distinguishing between molecules that differ in one or few mismatches.

It is also interesting to note that $N_{bins} = 200$ appears to be a reasonable choice for all classes for both labeling schemes, except S3, S4, and S10. While we have addressed the issues with (S3, S4) in the preceding paragraph, we hypothesize that the increased accuracy of S10 as N_{bins} increases is related to the low-dimensional overlap of S2, S8, and S10, as illustrated in Figure 2-4 and Figure 2-5. Stated differently, these three classes benefit from a sufficiently high number of dimensions (bins), which allows the classifier to narrow in on the fine differences between (S2, S8) and S10. In general, coarse binning has the effect of smoothing out these subtle differences, resulting in a drop in class accuracy.

2.2.5 Number of sample traces

Finally, we focus on performance analysis of baseline classifiers with respect to the number of traces used to compute a conductance histogram, *H*. Figure 2-12 (e) and (f) show the accuracy of the baseline classifiers w.r.t the number of traces used to compute a conductance histogram, *H*, corresponding to the two different target labeling schemes, maintaining $N_{bins} = 600$ and $\beta = 0.95$ (for detailed confusion matrices, see Appendix Figure A-3). In this case, we expect '*more is better*' from a performance perspective when it comes to choosing a proper value for *H*, since

histograms constructed from a sufficiently large number of traces should more closely approximate the underlying conductance distribution. This is corroborated in Figure 2-12 (e) and (f), with the only exception being S7 for TLS-2. For this class, accuracy shows a consistently decreasing trend, dropping from almost 100% to 98.9% as *H* is increased from 10 to 50. From our simulations, we find that the misclassification probabilities for S7 are 0.009 and 0.011 for *H* = 40 and 50 respectively, and in both cases, the probability that S7 is predicted to be S5 is 0.007, based on results averaged over 100 runs. Therefore, whenever S7 is misclassified, about 70% of the time, it is incorrectly predicted to be S5. At first glance, the reason for this exception might appear to be related to data scarcity since dataset S7 has 538 valid conductance traces to sample from at $\beta = 0.95$. With larger values of *H* such as 40 or 50, it is plausible that the corpus of training histograms for S7 is not diverse enough for the ensemble classifiers to properly learn to distinguish S7 from other classes in the database. However, if we observe the accuracy of S1, it does not exhibit a similar performance degradation as S7, despite the fact that only 528 valid traces remain at $\beta = 0.95$. An explanation for this contradictory behavior can be derived from a visual examination of the large sample conductance histograms shown in Figure 2-6. To the naked eye, the large sample histograms for S5 and S7 appear somewhat similar, while S1 stands out from all other datasets due to its pronounced peak at a conductance value of approximately $10^{-6}\ G_0$. For smaller values of *H*, we believe that the inherent small sample variability induces small differences (which can be viewed as additive noise) between S5 and S7 histograms, which aids with classifier generalization. Stated differently, we suspect that the XGboost classifier may be overfitting on S7 at larger values of *H*, which could be compounded by the data scarcity issue. On the other hand, we hypothesize that the acuity of the signature dominant peak at $10^{-5}\ G_0$ for S1 is relatively immune to the choice of *H*, which ensures that its classification accuracy remains consistently high across *H*.

While a higher value of *H* may be generally better from a performance perspective, it imposes an undue burden in practice since an SMBJ experiment needs to be conducted multiple times to make a highly accurate visual prediction. Specifically, if the probability of any experiment being valid (with molecular binding) is *p*, which of course depends on the R^2 test threshold parameter β, the average number of times an experiment needs to be conducted is H/p.

2.2.6 Analysis of secondary binary classifier for S3 and S4

In the above sections, we have established that a single XGboost classifier works phenomenally well in classifying the ten datasets into six (strand-based) classes or eight (strand, bias voltage-based) classes, with the exception of S3 and S4 which differ by a single mismatch. Nevertheless, based on our observations in conjunction with Figure 2-4, we recommend that experimental data on the same strand but with different bias voltages be coded as different classes. We have also demonstrated the challenges in designing a monolithic classifier with one set of values for (N_{bins}, β, *H*) which works equally well for all data classes. An alternate architectural option might be to employ a 2-stage tandem classifier where S3 and S4 are

combined as one target class, say superclass 3_4, in the first-stage primary classifier, followed by a second-stage binary classifier operating on S3 and S4 with a lower value of β and other parameters, suitably tuned for these classes. In this option, the second stage classifier is invoked if the prediction from the first stage indicates a 'class 3_4'. Another architectural option might be to design the primary classifier with 6 or 8 classes (S3 and S4 coded as different classes) and pair it with a secondary binary classifier operating on S3 and S4 for additional verification. If both classifiers predict the same class, that should obviously enhance the likelihood of correct identification. We believe that this tandem classifier approach, although motivated by our need to reject invalid experimental observations (driven by the R^2 test threshold, β), will be useful to distinguish between strands with similar conductance distributions, even if uncertainties associated with experimental conditions are rendered moot. In this section, we discuss the performance of a secondary binary classifier operating on S3 and S4 only.

Often, the performance of an unsupervised clustering algorithm on a labeled database is used as a benchmark for classifier performance. Toward that goal, we report the performance of a baseline spectral clustering algorithm on S3 and S4, using N_{bins} = 600 and H = 30. Given a set of n points, $\{x_i : 1 \leq i \leq n\}$, in d-dimensional space, the steps involved in a spectral clustering [123] approach are as follows: (i) construct a similarity graph on the n points, (ii) perform an eigen-decomposition of the Laplacian matrix of the similarity graph, and (iii) use a clustering method such as k-means to cluster the points using the p ($p \leq d$), eigenvectors of the graph Laplacian corresponding to the p smallest eigenvalues. The similarity between points x_i and x_j in the similarity graph is computed as follows:

$$s_{ij} = \exp\left(-\|x_i - x_j\|^2 / \sigma^2\right) \tag{2-2}$$

where σ is the scale parameter. Several methods exist for constructing the similarity graph; we use the nearest neighbor graph construct with a specified number of neighbors for each point. In addition, several forms of the Laplacian matrix are possible; we use the symmetric normalized Laplacian matrix defined in [124]. Figure 2-13 shows the class accuracies as a function of the number of neighbors, for p = 2, σ = 1, and β = 1, 0.95, and 0.87 (these simulations were conducted in MATLAB). Generally speaking, the clustering quality can be sensitive to the choice of σ. For our datasets, however, we found that the solution is robust to $\sigma \in [0.5, 2]$. Representative benchmark accuracy indices for S3 and S4 are (i) [82.9, 90.8] % for β = 1, (ii) [99.2, 91.3] % for β = 0.95, and (iii) [100, 99.1] % for β = 0.87, all numbers corresponding to 20-nearest neighbor similarity graphs. We observe that the spectral clustering method works remarkably well for both classes at β = 0.87. At β = 0.95, the accuracy of S3 is almost 8% better than the accuracy of S4, which, quite interestingly, is reversed at β = 1. While some degree of explanation of this behavior can be offered from a graph-cut viewpoint of the spectral clustering method, possibly applied to a low dimensional projection of the data, it is not necessary for the context of this thesis.

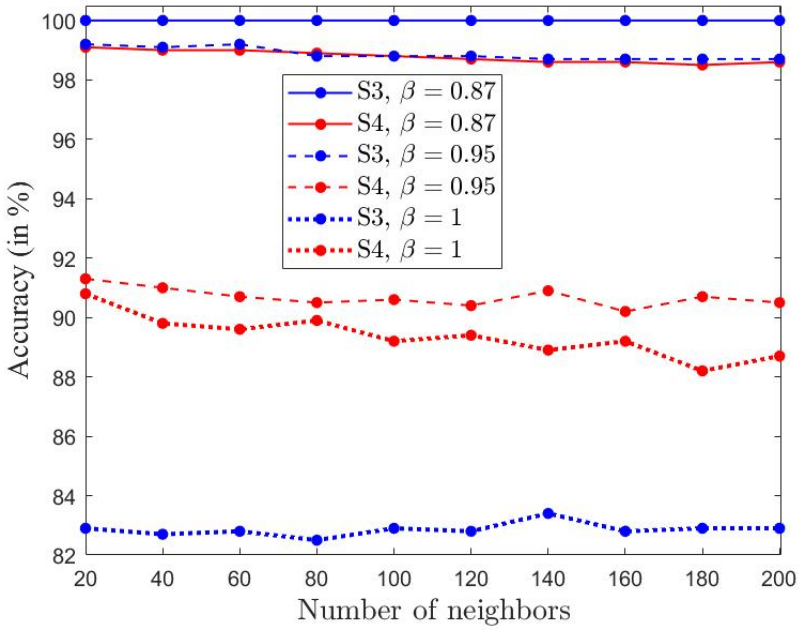


Figure 2-13. Clustering of S3 and S4 data. Class accuracies obtained from spectral clustering of S3 and S4 conductance histogram data (1000 per class) as a function of the number of neighbors used to construct a similarity graph on the data, using N_{bins} = 600 and H = 30.

We now discuss the performance of a secondary binary support vector machine (SVM) classifier trained with 700/300 training/testing histograms per class. Our choice of SVM as the classification algorithm was driven primarily by the clean separation which appears to exist between two-dimensional projections of the two classes at β = 0.87, as is evident from the *t*-SNE and classical multidimensional scaling (MDS) [120] visualizations in Figure 2-4 and Figure 2-5. Simulations were conducted in MATLAB using a vanilla SVM (linear kernel) with a regularization parameter of C = 1. The matrix of training histograms was standardized on a per-bin basis. While we did not find any evidence that standardization enhances XGboost's performance, the binary SVM classifier benefitted immensely from standardized data. Viewing the mismatched sample, S4, as the positive class and S3 as the negative class, training with unstandardized data resulted in substantially lower sensitivities but relatively high specificities, which is not desirable if detection of a mutant strain is the primary objective, even at the expense of a slightly higher false positive rate. Please see Figure 2-14 for details.

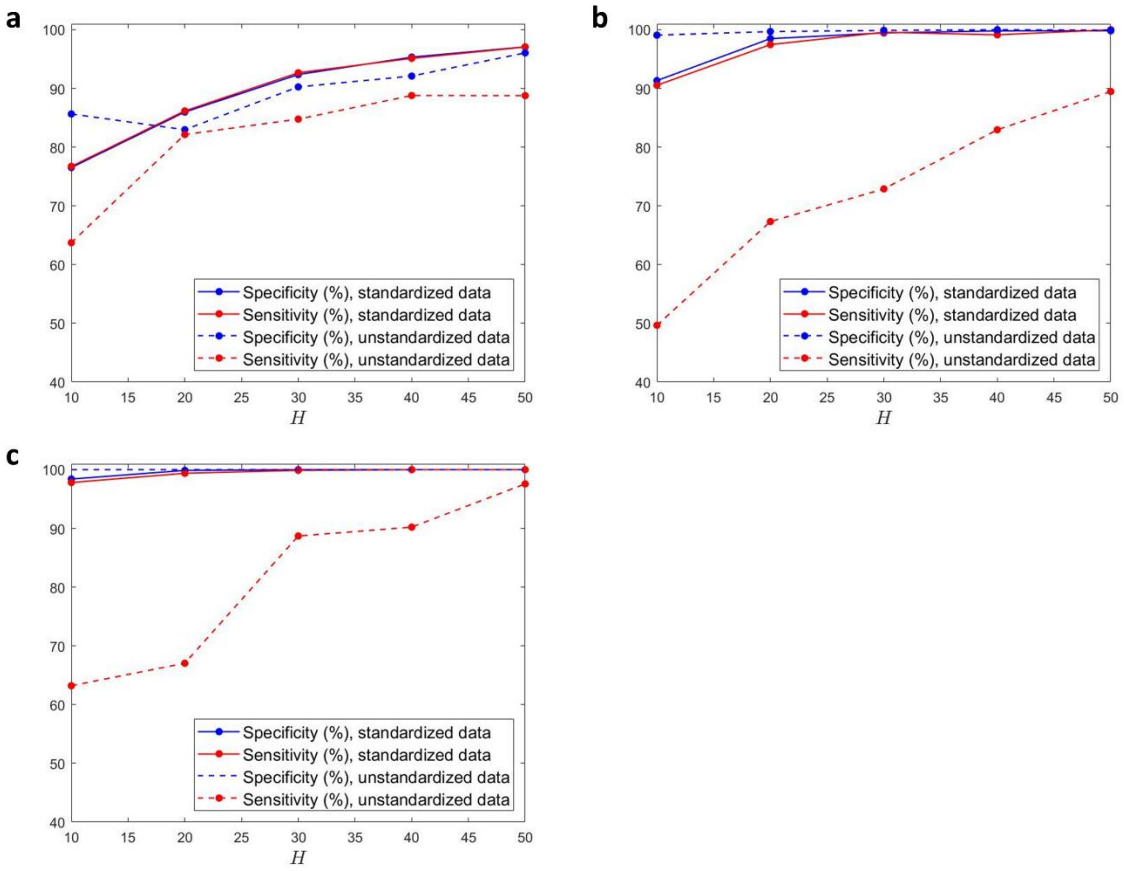


Figure 2-14. Binary SVM classifier on S3 and S4. We view the mismatched sample, S4, as the positive sample and S3 as the negative sample. This figure illustrates the impact of data standardization on the sensitivity and specificity of a binary vanilla SVM classifier (linear kernel with regularization parameter $C = 1$) operating on S3 and S4 with (a) $R^2 \leq \beta = 1$, (b) $R^2 \leq \beta = 0.95$, and (c) $R^2 \leq \beta = 0.87$. When the data is standardized, the classifier is naturally balanced which leads to approximately equal sensitivities and specificities. When the data is not standardized, we end up with a rather low sensitivity but high specificity classifier. While it is feasible to tune the classifier balance with unstandardized data, the primary objective of this figure is to illustrate the tremendous impact of standardization on the same baseline classifier.

Table 2-3 presents the classification accuracies with standardized data as a function of H and β, averaged over 100 runs, for $N_{bins} = 600$ and 60. First, we observe that class accuracies reach almost 100% for ($\beta = 0.87$, $H \geq 30$) and coarse binned histograms work almost as well as fine-grained histograms. The only cases where coarse histograms are decidedly better than fine histograms are when ($\beta = 1$, $H \leq 20$), but these aren't of practical interest due to rather low-class accuracies. Second, the class accuracies drop significantly as β increases, which is consistent with the behavior of the primary XGboost classifier. At $H = 30$, the drop is almost 7-8%, comparable to what we saw in Figure 2-12 (a) and (b). Third, at baseline values of ($\beta = 0.95$, $N_{bins} = 600$, $H = 30$), class accuracies are [99.45, 99.57] %, which are almost 2.5-3% better than

the accuracies indicated in Figure 2-8 for the primary classifier. This demonstrates the benefit of using a tandem classifier approach compared to a monolithic model.

Table 2-3. Class accuracies in percentages for a secondary binary SVM classifier operating on S3 and S4*

H	$\beta = 1$		$\beta = 0.95$		$\beta = 0.87$	
	S3	S4	S3	S4	S3	S4
10	76.53	76.73	91.33	90.54	98.41	97.82
20	85.97	86.17	98.48	97.46	99.88	99.39
30	92.36	92.67	99.45	99.57	100.00	99.89
40	95.33	95.12	99.80	99.13	100.00	100.00
50	97.08	97.05	99.82	99.99	100.00	100.00

H	$\beta = 1$		$\beta = 0.95$		$\beta = 0.87$	
	S3	S4	S3	S4	S3	S4
10	81.34	78.03	91.57	89.52	96.07	94.76
20	89.69	88.78	96.54	95.94	98.53	98.56
30	92.87	92.53	98.42	98.23	99.60	99.65
40	95.03	95.12	98.86	98.94	100.00	99.96
50	96.43	96.69	99.42	99.36	100.00	100.00

* The left and right tables correspond to N_{bins} = 600 and N_{bins} = 60 respectively. The accented rows and columns correspond to baseline parameter values (β = 0.95, N_{bins} = 600, H = 30). For comparison, the (S3,S4) class accuracies obtained from spectral clustering with (N_{bins} = 600, H = 30) are (i) [82.9, 90.8] % for β = 1, (ii) [99.2, 91.3] % for β = 0.95, and (iii) [100, 99.1] % for β = 0.87.

2.3 Summary

Identification of genetic material from signatures in current traces has been pursued for over a decade. The conductance histograms from experimentally measured currents are intrinsically noisy, with values ranging over three orders of magnitude. Further, physics-based theory and modeling approaches have so far been ineffective in capturing the noisy nature of experimental data with enough accuracy to reveal sequence information. This is because the number of atoms involved is extremely large, the environment fluctuates, and there are a large number of DNA-contact configurations, which makes it impossible to model the system in a realistic manner. As a result, identifying genetic material from current traces has remained a challenge. In this work, we have demonstrated that an ML-based approach using experimentally obtained current traces is extremely successful in identifying DNA strands. Conductance traces were sampled randomly (H at a time) to construct the conductance histograms and an XGboost classifier was trained on the histograms. The computational time for strand identification from raw experimental data using a trained model is as small as 28 μs, which makes our approach suitable for real-time implementation. For molecules that are sufficiently different structurally, we obtain an accuracy of 99.5% with $\beta \geq 0.95$ and H as small as 20 – 30. For DNA-RNA hybrids from E. coli with just a single base pair mismatch, our methods can achieve similar accuracy, but with a more stringent set of parameters, β = 0.90 and H = 50. In general, we observe that a monolithic classifier model trained on multiple data classes may not provide comparably high accuracy for genetic samples which differ by a single mismatch, especially if H is kept reasonably low in the interest of rapid detection and classification in practice. Our investigations reveal that a tandem classifier approach, where the first stage is a multiclass classifier and the second stage is a binary classifier operating exclusively on molecules with single base pair mismatches, can be an attractive

architectural proposition for boosting the accuracies of such samples to around 99.5% with a reasonable $H = 30$. Overall, our approach shows tremendous potential for accurate, fast, cheap, and amplification-free DNA strand identification. Extremely short computational times, along with demonstrated high accuracies, make single-molecule sensing using the current traces possible in real time and establishes it as a feasible candidate for diverse time-critical sequence identification applications, including clinical diagnostics.

3 Utilize 2D Electrical Conductance Probability Distributions from Mixed Data Sets

In this thesis, 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 chapter (published in reference [91]) was collected using the single molecule break junction (SMBJ) approach under a range of experimental parameters: applied voltage bias, current amplifier sensitivity, distance between electrodes, and ramp rate (explained in the Methods Section 3.1). As a benchmarking step, we first applied the method we proposed in [38] to the data, which yielded low accuracies (83.96%). Then we develop a new methodology based on (i) two-dimensional (2D) conductance probability distributions instead of one-dimensional (1D) distributions and (ii) averaged (over the experimental parameters) conductance distributions instead of non-averaged distributions. We trained XGBoost classifiers on 1D distributions and convolutional neural networks paired with XGBoost classifiers on 2D distributions. Our primary research objective was to explore whether any or both of the above variations would better capture the electronic fingerprints of the sequences, which could translate to improved classifier models. Our secondary objectives were: (i) how many experimental passes are needed to construct reliable conductance distributions? and (ii) what role does the value of the applied voltage between the electrodes play, if any? We provide an in-depth analysis of our investigations in the Results Section 3.2. Our investigations 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) 20 experimental passes are adequate to yield classifiers which exceed 90% accuracy for six of the ten sequences considered in this chapter, but at least 50 passes are needed for four other sequences, (iv) data collected from experiments conducted at lower voltage biases tend to produce better classifiers, and, (v) higher biases can produce contradictory behavior in classifier accuracy, depending on the sequence type.

3.1 Methodology

3.1.1 Conductance datasets

The data for the conductance is from reference [91], which used the SMBJ method for data collection. A simplified overview of this method is shown in Figure 3-1. The experimental apparatus consists of two electrodes, the metal substrate, and the tip. Initially, a layer of self-assembled DNA molecules is fabricated on the substrate and a fixed voltage bias is applied between the tip and substrate. In the next step, the metal tip approaches the surface and contacts the substrate resulting in a large current. The tip is then slowly retracted, with the distance between the tip and substrate increasing and the current is continuously monitored as a function of time. In many measurements, a DNA molecule is attached between the tip and substrate, which yields a current versus time trace that is nonexponential (has plateaus) - these are the

measurements that are of interest. There are other spurious measurements where there is no DNA between the tip and substrate, and these result in an exponential decrease in current versus time. Finally, with further retraction of the tip, the DNA loses contact with the substrate as shown in Figure 3-1, at which point, the current falls to zero (the noise floor of the apparatus). The method is intrinsically noisy because the atomistic details of the contact between the DNA and metal can have multiple configurations, see Figure 3-2. Further, the immediate environment of the DNA as the tip is retracted can also change between measurements. Yet several studies have obtained insights on molecular scale transport using this method that was pioneered in the 1980s [125]. The ramp rate in Figure 3-1 determines how fast the distance between the tip and substrate increases with time when the tip is retracted. In reference [91], the experimental datasets used a voltage bias from the set $\{0.03, 0.05, 0.10, 0.15, 0.20, 0.30\}\ V$ and a ramp rate from the set $\{3, 5, 10, 20\}\ V/s$. Further, both the lower and upper limit of the current range is set by the amplifier. For the current amplifiers with 1 and 10 nA/V sensitivities, the current range is [1 pA – 10 nA] and [10 pA – 100 nA] respectively [91]. The experimentally measured current is sensitive to 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.

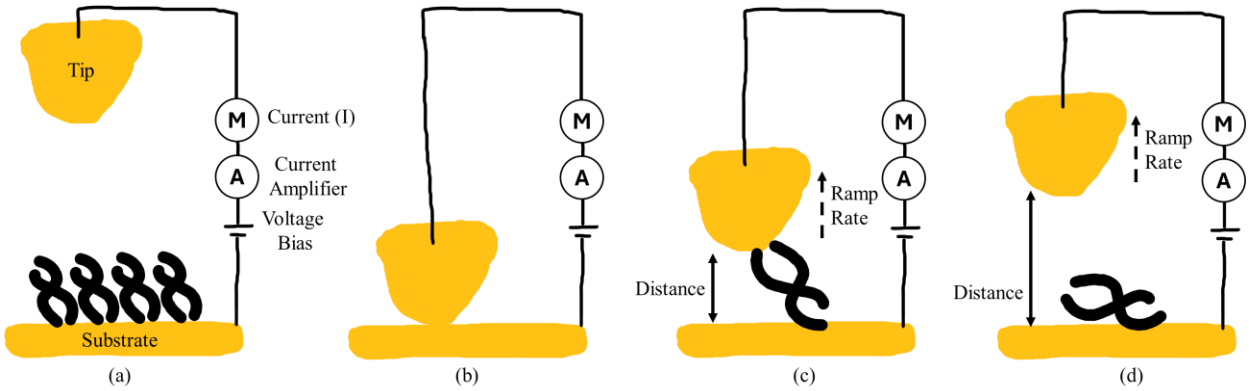


Figure 3-1**.** (a) A self-assembled layer of DNA, (b) The tip initially makes contact with the substrate when the current is large. (c) As the tip is retracted current flows from the tip to the substrate via the DNA. Only one DNA strand is shown for clarity. (d) With further retraction of the tip, electrical contact is broken, and the current becomes zero. The current is continuously monitored during the tip retraction process.

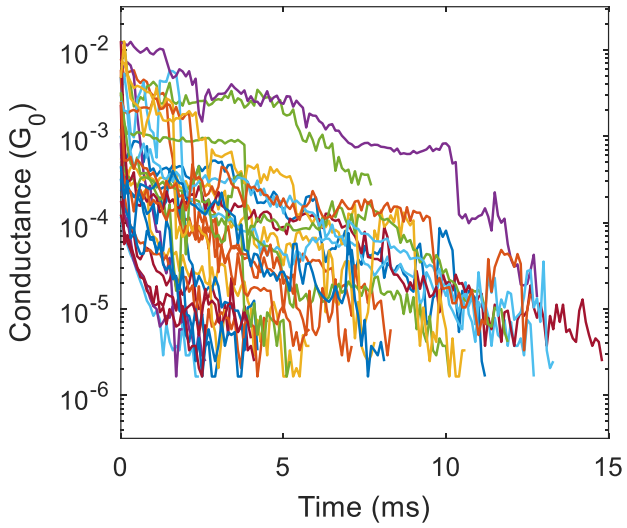


Figure 3-2. Twenty conductance traces that randomly sampled from one dataset, which comes from same DNA molecule and same experimental setup.

The dataset used in this work 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 published in [91], and detailed in Table 3-1. These variants are the Alpha variant B.1.1.7, Beta variant B.1.351, and the Delta variant B.1.167 [101]. 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 denoted by Alpha_MM2, Beta_MM2, and Delta_MM2 are hypothetical sequences that could produce the same virus protein as the PM sequences [91]. Table 3-1 shows the exact mutation locations for the ten sequences considered along with the experimental parameters for measurement, which resulted in a total of 39 datasets.

Data acquisition. We relied on SMBJ measurements to capture the conductance characteristics of the DNA sequences. Generally, these experiments (preparing the RNA-DNA hybrids and obtaining conductance traces) take around 4 hours to generate 5,000 conductance traces. We obtained all SMBJ experimental data from [91]. For a detailed explanation and experimental setup, refer to [16], [38]. The experimentally obtained SMBJ current traces are not uniform in duration. This limits the applicability of machine learning classification methods like XGBoost. Therefore, a series of pre-processing steps are used to convert the SMBJ current traces to probability histograms, as explained in the following Section 3.1.2 and [38]. These histograms have a uniform number of logarithmically spaced bins, introducing uniformity to the current traces. Additionally, a single probability histogram is constructed from multiple traces, thus averaging out the presence of random, high-frequency noise.

Table 3-1. Summary of the 39 COVID-19 datasets used.

Label	Sequence	Voltage Bias	Current Amplifier	Ramp Rate
E1	***Alpha* variant, *mismatch 1***	0.10 V	1 nA/V	5 V/s
E2		0.15 V	10 nA/V	3 V/s
E3	3'- GGG TGA ATA CCA -5'	0.10 V	10 nA/V	10 V/s
E4	5'- CCC ACU **A**AU GGU -3'	0.15 V	10 nA/V	10 V/s
E5		0.20 V	10 nA/V	10 V/s
E6	***Alpha* variant, *mismatch 2***	0.15 V	1 nA/V	3 V/s
E7		0.10 V	1 nA/V	10 V/s
E8	3'- GGG TGA ATA CCA -5'	0.10 V	10 nA/V	10 V/s
E9	5'- CCC ACU UA**C** GGU -3'	0.15 V	10 nA/V	10 V/s
E10		0.20 V	10 nA/V	20 V/s
E11	***Alpha* variant, *p*erfect *m*atch**	0.05 V	1 nA/V	3 V/s
E12		0.10 V	1 nA/V	3 V/s
E13	3'- GGG TGA ATA CCA -5'	0.10 V	1 nA/V	20 V/s
E14	5'- CCC ACU UAU GGU -3'	0.10 V	10 nA/V	10 V/s
E15		0.20 V	10 nA/V	10 V/s
E16	***Beta variant, mismatch 1***	0.10 V	10 nA/V	20 V/s
E17	3'- CCA CAA TTT CCA -5'	0.05 V	10 nA/V	20 V/s
E18	5'-GGU GUU **G**AA GGU-3'	0.10 V	10 nA/V	20 V/s
E19	***Beta variant, mismatch 2***	0.30 V	10 nA/V	20 V/s
E20	3'- CCA CAA TTT CCA -5'	0.30 V	10 nA/V	20 V/s
E21	5'-GGU GUU AA**G** GGU-3'	0.10 V	10 nA/V	20 V/s
E22	***Beta variant, mismatch 3***	0.05 V	10 nA/V	20 V/s
E23	3'- CCA CAA TTT CCA -5'	0.10 V	10 nA/V	20 V/s
E24	5'-**A**GU GUU AAA GGU-3'	0.10 V	10 nA/V	20 V/s
E25	***Beta variant, perfect match***	0.05 V	1 nA/V	20 V/s
E26	3'- CCA CAA TTT CCA -5'	0.10 V	1 nA/V	20 V/s
E27	5'-GGU GUU AAA GGU-3'	0.10 V	10 nA/V	20 V/s
E51	***Delta variant, mismatch 1***	0.03 V	10 nA/V	20 V/s
E52	3'-TCG TTT GGA ACA-5'	0.03 V	10 nA/V	20 V/s
E53	5'-AGC A**C**A CCU UGU-3'	0.10 V	10 nA/V	20 V/s
E54		0.10 V	10 nA/V	20 V/s
E55	***Delta variant, mismatch 2***	0.03 V	10 nA/V	20 V/s
E56	3'-TCG TTT GGA ACA-5'	0.10 V	10 nA/V	20 V/s
E57	5'-AGC AA**G** CCU UGU-3'	0.10 V	10 nA/V	20 V/s
E58		0.20 V	10 nA/V	20 V/s
E59	***Delta variant,* perfect *m*atch**	0.03 V	10 nA/V	20 V/s
E60	3'-TCG TTT GGA ACA-5'	0.03 V	10 nA/V	20 V/s
E61	5'-AGC AAA CCU UGU-3'	0.10 V	10 nA/V	20 V/s
E62		0.20 V	10 nA/V	20 V/s

Data visualization. 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 [91], as detailed in Table 3-1. Appendix Figure B-1, Appendix Figure B-2, Appendix Figure B-3 show the empirical *large sample* 1D conductance histograms for our 39 datasets (see Table 3-1) using an R^2 test threshold of $\beta = 0.95$. We refer to these histograms as large sample histograms since these are based on thousands of conductance traces for each dataset. Each row in Appendix Figure B-1, Appendix Figure B-2, Appendix Figure B-3 represents a specific sequence (e.g., Alpha_MM1, Beta_PM, etc.). The variability in the shapes of the histograms shown in each row can be attributed to the inherent stochasticity of the SMBJ experimental procedure and/or variations in experimental parameters such as voltage bias and ramp rate. Similarly, Appendix Figure B-4, Appendix Figure B-5, Appendix Figure B-6 show the empirical large sample 2D conductance histograms for our 39 datasets using an R^2 test threshold of $\beta = 0.95$.

3.1.2 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. Figure 3-3 summarizes the sequence of pre-processing steps. First, the current traces were clipped to the range [1 pA – 10 nA] or [10 pA – 100 nA] depending on the amplifier used. 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 on every run of the experimental procedure is not guaranteed, the spurious current traces which are predominantly exponentially decaying (in contrast, traces with molecular binding exhibit plateaus) are removed using a R^2 test (invalid data). Specifically, if the computed R^2 statistic is greater than some chosen threshold, β, we reject the trace as 'invalid'); otherwise, the trace is accepted. 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 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 $\beta = 0.95$.

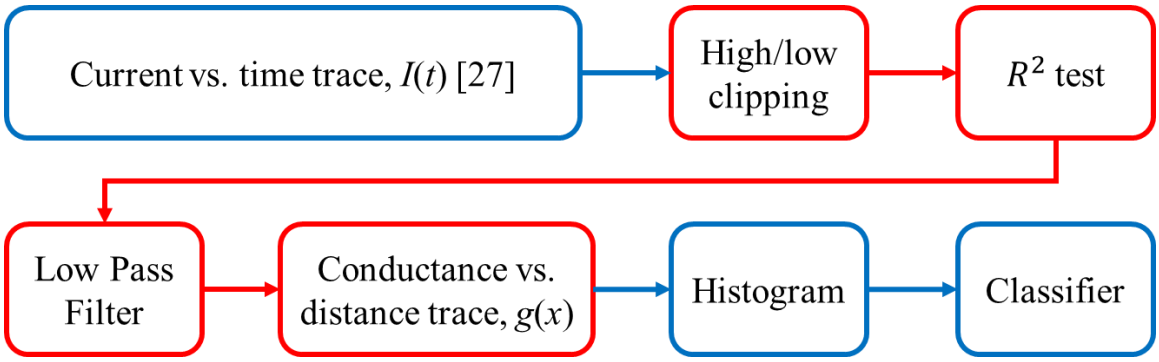


Figure 3-3. Sequence of pre-processing steps (shown within red boxes) for converting raw current traces to conductance traces.

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:

$$Conductance\ (G_0) = \frac{Current}{V_{bias}} \tag{3-1}$$

where $Current$ is in units of 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 $g(x)$ traces using the eq. below:

$$x = Distance\ (nm) = 4 \cdot t \cdot Ramp\ Rate \tag{3-2}$$

where each time step is 10^{-4} s, corresponding to a sampling frequency of 10 kHz. Ramp rate, in V/s, is the voltage applied to control the rate at which the distance between the two electrodes changes. The factor '4' in eq. (*3-2*) accounts for the fact that the velocity of the top electrode is proportional to the piezo displacement, which is 4 nm/V. The distance between the electrodes is in units of nm, and most measurements have a distance value of less than 0.4 nm. Therefore, we choose [0−0.4] nm as the range of distance values for creating histograms.

3.1.3 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. Figure 3-4 shows sample 2D and 1D histograms constructed from a single conductance trace along with a schematic of the experimental setup and the parameters (bias, ramp rate, current amplifier

sensitivity). 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, see Figure 3-4) 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 the following sections, we have experimented with two different sampling methods for the conductance traces, resulting in either a *conductance histogram* or an *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 3-1). 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*.

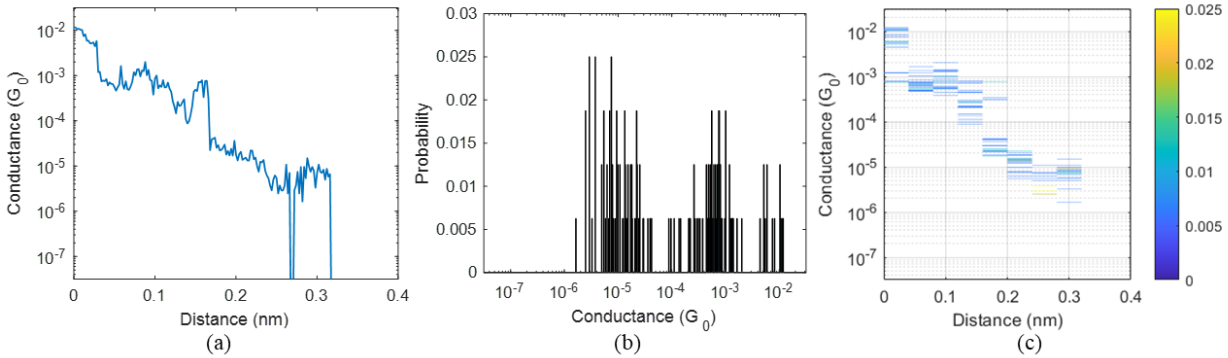


Figure 3-4. (a) Schematic of the SMBJ experimental method, (b) a single conductance versus distance trace, (c) derived 1D conductance histogram, and (d) derived 2D conductance histogram.

Irrespective of the method used to sample H conductance traces, we can construct either a 1D histogram or a 2D histogram. We now elaborate on the procedures. As recorded, each conductance trace is a function of time, which can be recalibrated in terms of the distance between the two contacts (as time increases, so does the inter-electrode distance), as discussed in Section 3.1.2. Since the majority value of the inter-electrode distance is approximately less than 0.4 nm, we divide the range of distance values, $[0-0.4]$ nm, into a number of distance bins, which we denote by $N_{bins-Distance}$. Similarly, the range of conductance values is divided into a number of conductance bins, which we denote 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. The 2D histograms, therefore, provide additional information regarding the inter-electrode distance where conductance change occurs, which we have found can be critical in successfully distinguishing between certain SARS-CoV-2 variants. By summing over the distance axis, we can marginalize the 2D histogram, yielding a 1D conductance histogram. Please refer to Appendix Figure B-1, Appendix Figure B-2, Appendix Figure B-3, and Appendix Figure B-4, Appendix Figure B-5, Appendix Figure B-6 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.

Next, we construct 1000 histograms for each sequence, 70% of which are used for classifier training and validation, while the rest (30%) are used for testing.

3.1.4 Machine learning architecture

We used the Tensorflow implementation of CNN, with a learning rate based on the Adam optimizer [126]. We employ categorical cross entropy for the multiclass classification. Figure 3-6 shows the architecture of our stacked CNN feature extractor and XGBoost model. The inputs to the XGBoost model are the outputs of the flattening layer from a CNN model trained with a fully connected classifier. Adopting an ensemble classifier model (XGBoost) instead of a traditional fully connected classifier results in an enhancement of more than 5% classifier accuracy. Figure 3-5 shows the comparison between the CNN-only approach and the CNN + XGBoost approach. We reckon this is due to the inherent stochasticity of the experimental data, which introduces a variance in the conductance probability distributions. Ensemble learning methods, in general, are better suited to data that is noisy and exhibit high variance. All codes/data used for generating the results are available as a Python package [127].

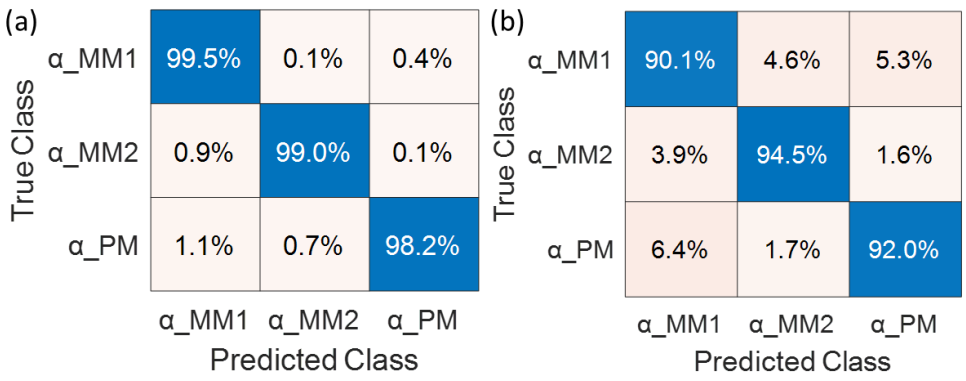


Figure 3-5. Confusion matrices for Approach A4 using (a) CNN + XGBoost and (b) CNN only as a machine learning algorithm for classification.

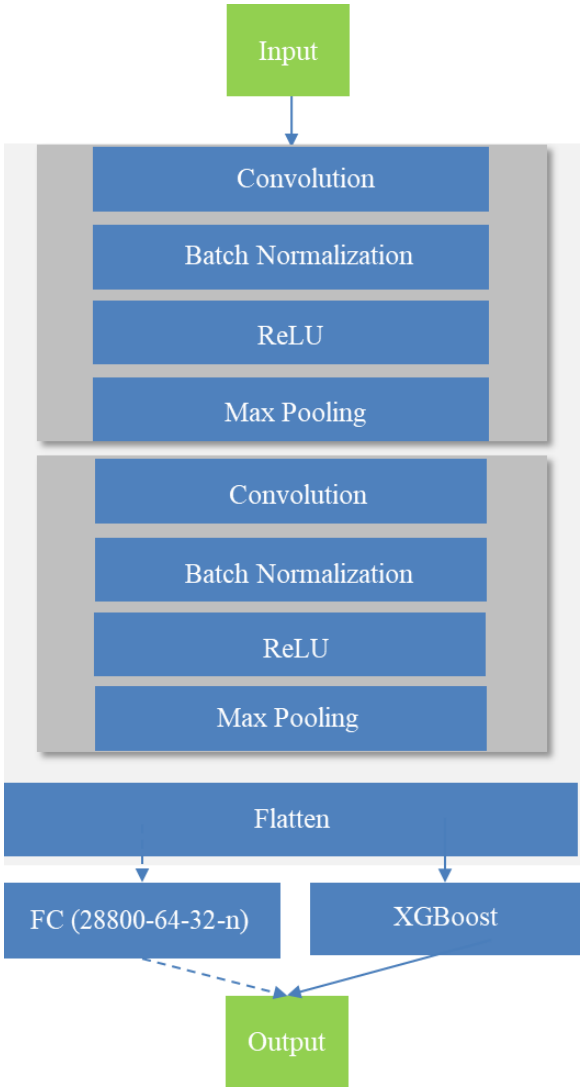


Figure 3-6. Sequence of components in the stacked CNN and XGBoost model. FC stands for fully connected layer.

3.2 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 3-2 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 that 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 3-2. In general, the performance of the classifier depends on the choice of the parameters $\beta, H,$ and N_{bins}, of which the sample size parameter H has the most pronounced effect.

3.2.1 Performance of baseline classifier

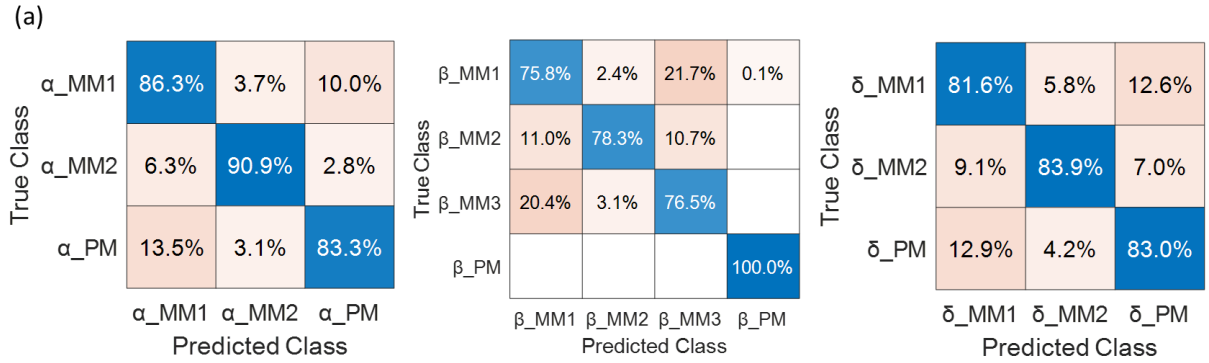
In this section, we discuss the performance of a *baseline classifier* by comparing in detail the differences between approaches A1 (worst performing) and A4 (best performing) indicated in Table 3-2. We observe that for small sample sizes such as $H = 10$, approach A4, which utilizes 2D average conductance 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 3-2.

Table 3-2. Summary of four different input data representations, ML algorithms used, and corresponding classification accuracies.

Approach	Histogram dimensionality	Average conductance histogram?	ML algorithm	Average accuracy	
				$H = 30$ (baseline)	$H = 10$
A1 [38]	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%**

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.

Figure 3-7 shows the confusion matrices for the Alpha, Beta, and Delta baseline classifiers; panels (a) correspond to Approach A1 [38], while panels (d) correspond to Approach A4. Please refer to Figure 3-7 for the confusion matrices of all four approaches. Comparing Figure 3-7 (a) and (d), we see that 1D histograms work reasonably well (≈ 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 Figure 3-7 (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 Appendix Figure B-2. 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 Figure 3-7 (c) and (f), we observe that 2D average histograms provide an average accuracy boost of approximately 4-10% for the Delta classes.



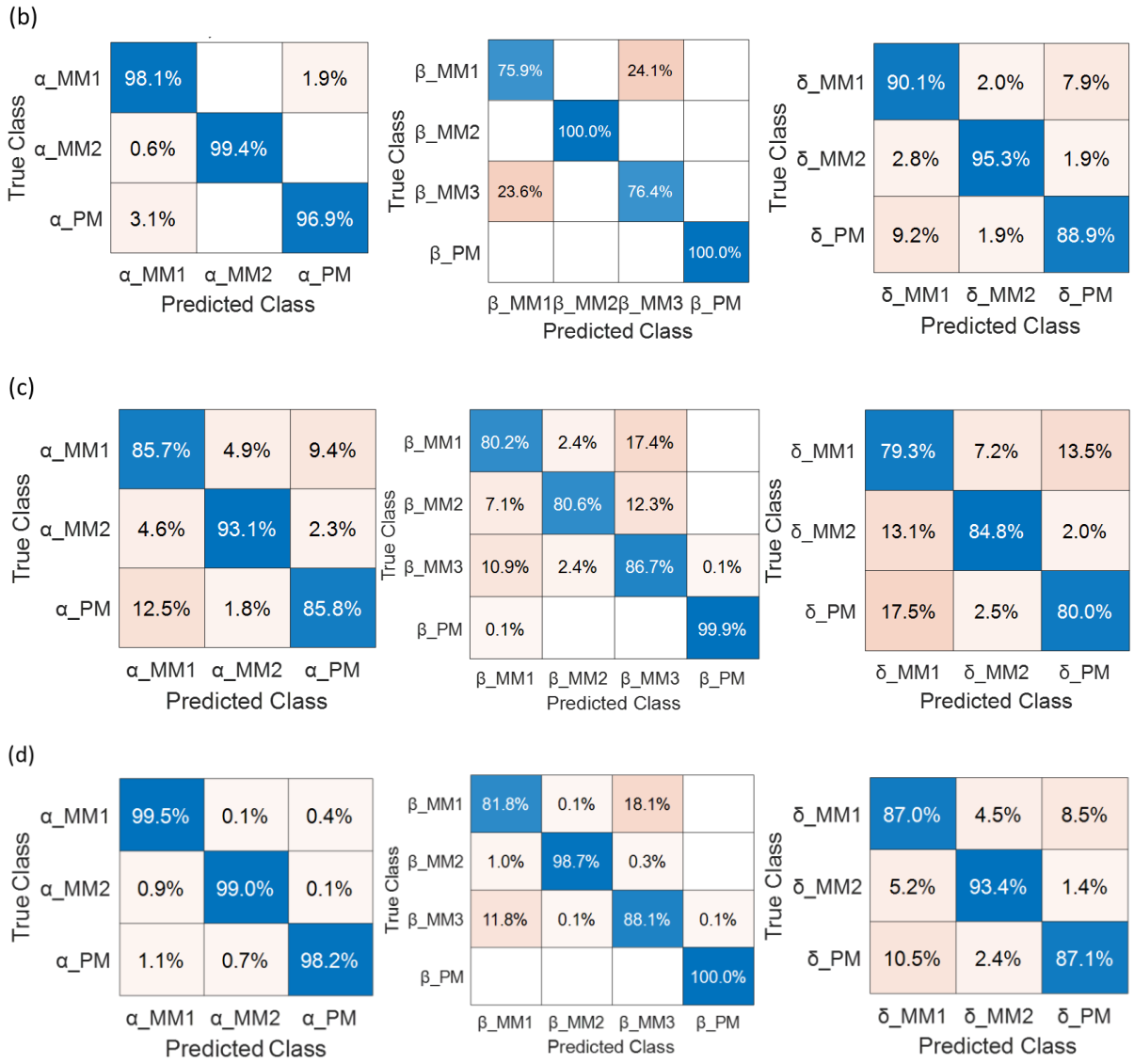


Figure 3-7. Confusion matrices for baseline classifiers: (a) Approach A1, (b) Approach A2, (c) Approach A3, and (d) Approach A4. The differences between these four approaches are summarized in Table 3-2.

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

Figure 3-8 shows the overall classification accuracy for all four approaches indicated in Table 3-2 with respect to the sample size, H, while keeping $\beta = 0.95$, $N_{bins} = 600$, and $N_{bins-Distance} = 10$. A similar plot over an extended range of H, $H \in [10,100]$, appears in Figure 3-9. The reason why we study all four approaches in this section is that Figure 3-8 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 Figure 3-8, 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.

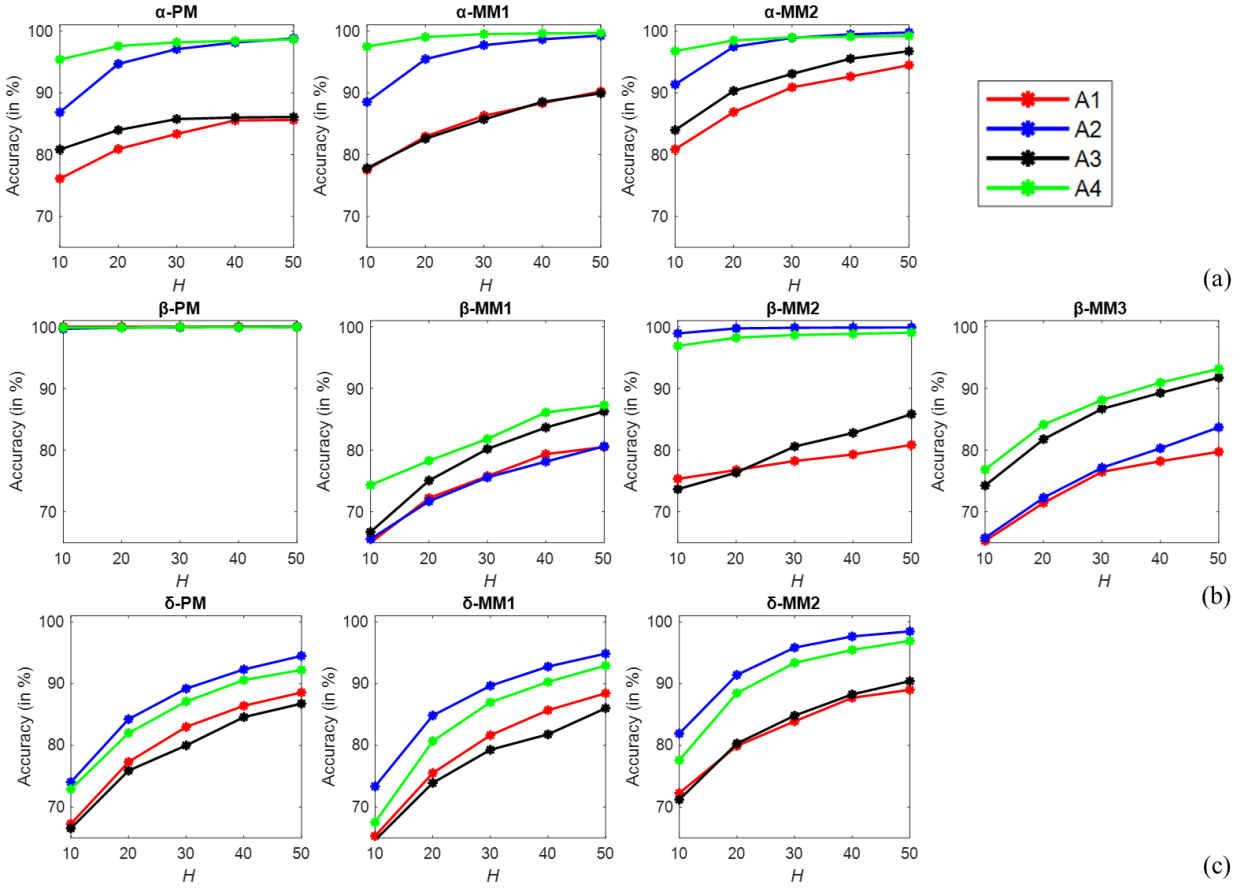


Figure 3-8. Performance analysis of baseline classifiers with respect to H for the four approaches indicated in Table 3-2. (a) Alpha sequences, (b) Beta sequences, and (c) Delta sequences. The same color scheme has been used to distinguish between the four approaches.

From the bottom row in Figure 3-8, 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 Figure 3-8, 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 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/tip distance indexed conductance distribution, but not all, is not understood at this point and is subject to further theoretical investigation.

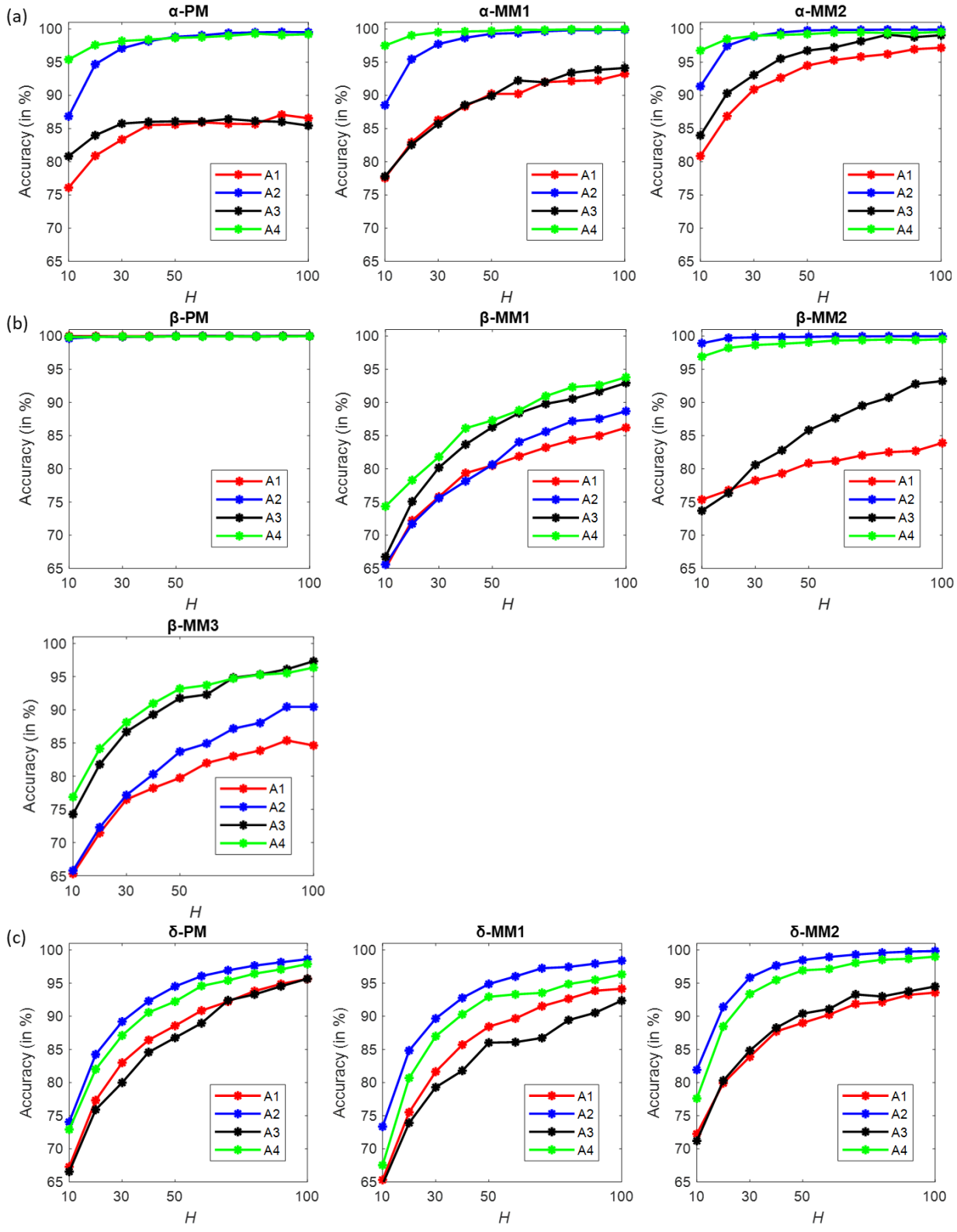


Figure 3-9. Performance analysis of baseline classifiers with respect to *H* for the four approaches.

3.2.3 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 *SMBJ 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 3-1 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 3-1 based on these two parameters. Subsequently, we combined datasets that 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 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 3-2) 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 3-3 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 Figure 3-10, Figure 3-11, and Figure 3-12. Results for classifiers trained on histograms with $H = 30$ are shown in Table 3-4.

Table 3-3. 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 $10\ V/s$.

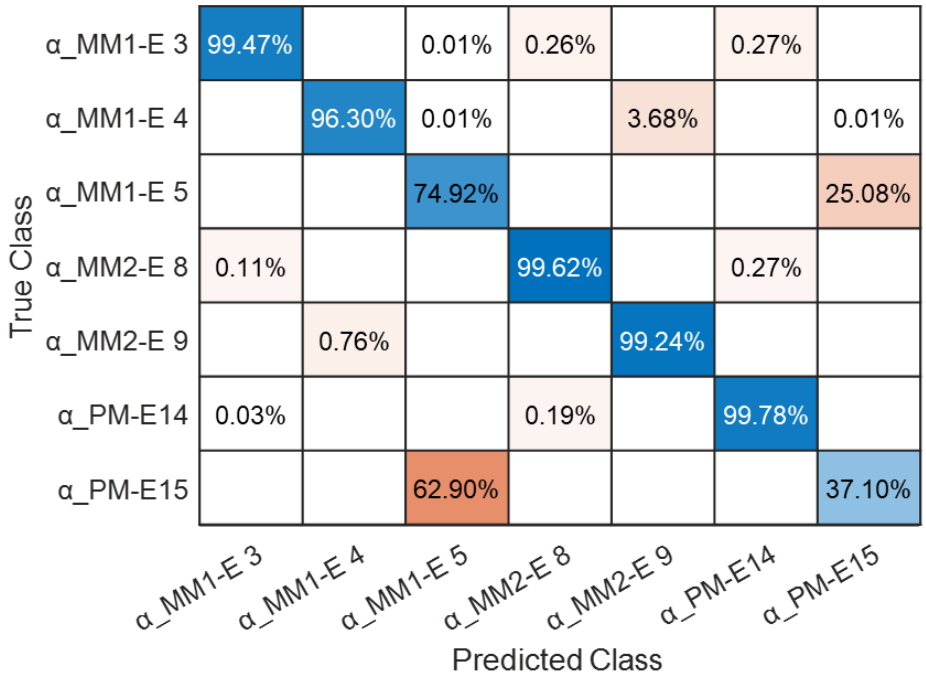


Figure 3-10. Confusion matrix of 7-class classifier for Alpha variant (approach A3, $H = 100$).

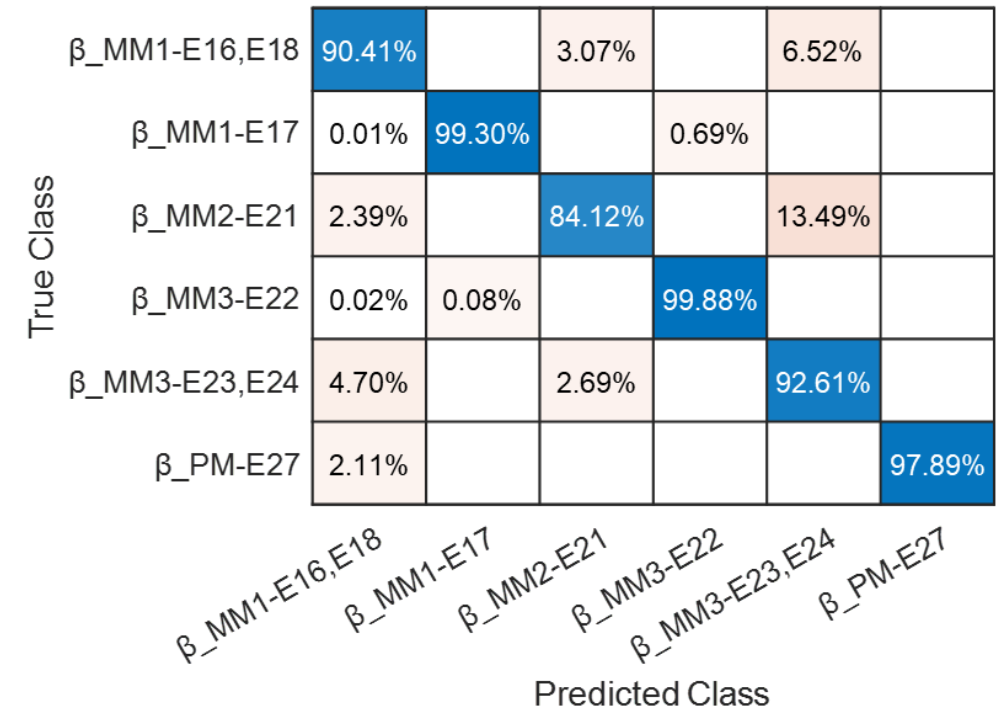


Figure 3-11. Confusion matrix of 6-class classifier for Beta variant (approach A3, $H = 100$).

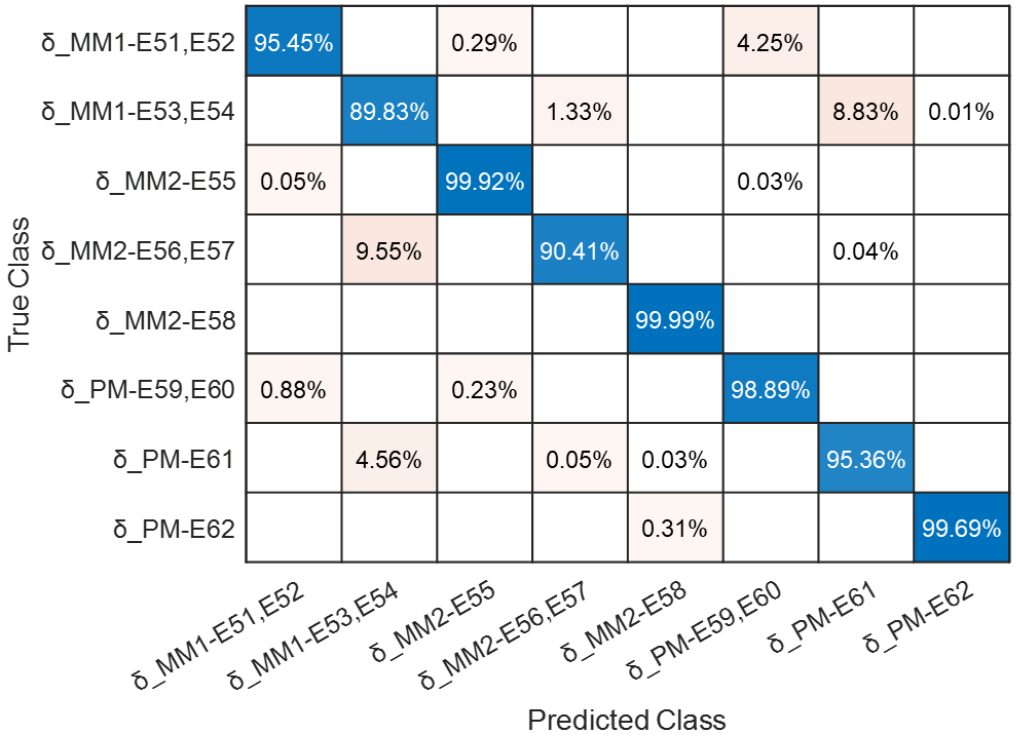


Figure 3-12. Confusion matrix of 8-class classifier for Delta variant (approach A3, $H = 100$).

A general observation from Table 3-3 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 3.2.4 (see Figure 3-13 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.

Table 3-4 shows the classifier accuracies for Alpha, Beta, and Delta variants as a function of applied bias using approach A3 and $H = 30$. The trend observed from Table 3-3 (lower biases result in higher accuracies) holds when classifiers are trained with $H = 30$ histograms, although the impact is somewhat obfuscated by sampling-induced uncertainties.

Table 3-4. 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	90.03 %	E4	83.49 %	E5	62.70 %
	MM2	E8	90.93 %	E9	89.43 %		
	PM	E14	91.80 %			E15	44.29 %
Mean accuracy			90.92 %		86.46 %		53.50 %

Voltage Bias		0.05 V		0.10 V	
Beta variant	MM1	E17	93.66 %	E16 ∪ E18	72.60 %
	MM2			E21	60.97 %
	MM3	E22	97.46 %	E23 ∪ E24	76.13 %
	PM			E27	99.99 %
Mean accuracy			95.56 %		77.42 %

Voltage Bias		0.03 V		0.10 V		0.20 V	
Delta variant	MM1	E51 ∪ E52	81.65 %	E53 ∪ E54	66.27 %		
	MM2	E55	96.61 %	E56 ∪ E57	77.91 %	E58	99.64 %
	PM	E59 ∪ E60	88.72 %	E61	82.12 %	E62	99.44 %
Mean accuracy			89.00 %		75.43 %		99.54 %

All histograms generated using $H = 30$. 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.

3.2.4 Validation by Jensen-Shannon metric

In order to better understand the contradictory behavior exhibited by the Alpha and Delta variants at a bias of 0.20 V (see Table 3-3), we first observe from the confusion matrix shown in Figure 3-10 that Alpha variant E5 (Alpha_MM1) is most often confused with E15 (Alpha_PM). Specifically, the probability that E5 is misclassified as E15 is 0.412, and the probability that E15 is misclassified as E5 is 0.584. Clearly, such high misclassification rates should imply a high degree of similarity between the histograms of E5 and E15. In order to verify this conjecture, we computed the pairwise Jensen-Shannon (J-S) distances[128] between the E5 and E15 histograms (normalized as probability distributions) and derived the cumulative distribution function (CDF) of J-S distances. Given two discrete probability distributions p_x and q_x on some random variable x, let m_x denote the pointwise mean of p_x and q_x, i.e., $m_x = \frac{p_x + q_x}{2}$. Then, the Jensen-Shannon (J-S) distance between p_x and q_x is defined as:

$$\text{J-S distance} = \sqrt{\frac{D(p_x \parallel m_x) + D(q_x \parallel m_x)}{2}} \tag{3-3}$$

where $D(p_x \parallel m_x)$ denotes the Kullback–Leibler divergence (KLD) of m_x from p_x, defined as follows:

$$D(p_x \parallel m_x) = \sum_i p_x(i) \log\left(\frac{p_x(i)}{m_x(i)}\right) \tag{3-4}$$

While the KLD is not necessarily symmetric, the Jensen-Shannon distance is. In the context of our data, let p_x denote a (1D) conductance distribution for the E5 class and q_x denote a (1D) conductance distribution for the E15 class. Suppose we have 1000 conductance representations for each class. We compute the J-S distances between each pair $\{p_x(m), q_x(n)\colon 1 \leq m, n \leq 1000\}$ and then plot the cumulative distribution function (CDF) of the computed J-S distances. This CDF plot is shown in Figure 3-13(a). Since the entire probability mass of J-S distance is concentrated within the band $[0.120, 0.175]$, we can infer that there is substantial similarity between the E5 and E15 histograms, which helps explain the high misclassification rates.

In contrast, we observe from the confusion matrix in Figure 3-12 that Delta variant E58 (Delta_MM2) is very occasionally confused with E56 ∪ E57 (Delta_MM2), which occurs with a probability of 0.00027. Similarly, Delta variant E62 (Delta_PM) is very occasionally confused with E58 (Delta_MM2) or E61 (Delta_PM), and each occurs with a probability of 0.00013. These extremely small misclassification rates should imply a low degree of similarity between the histograms of (i) E58 and E56 ∪ E57, (ii) E62 and E58, and (iii) E62 and E61. This conjecture can be verified by examining the corresponding CDF plots of J-S distances shown in Figure 3-13(b). We observe that the CDF plots in this case have shifted to the right and the probability mass of J-S distance is almost entirely concentrated within the band $[0.290, 0.485]$. Since the trailing edge of this distance band has shifted to the right compared to what we observe

for the Alpha variant, we can infer that the degree of similarity between the "problematic" Delta classes is relatively small, which helps explain the small misclassification rates for E58 and E62.

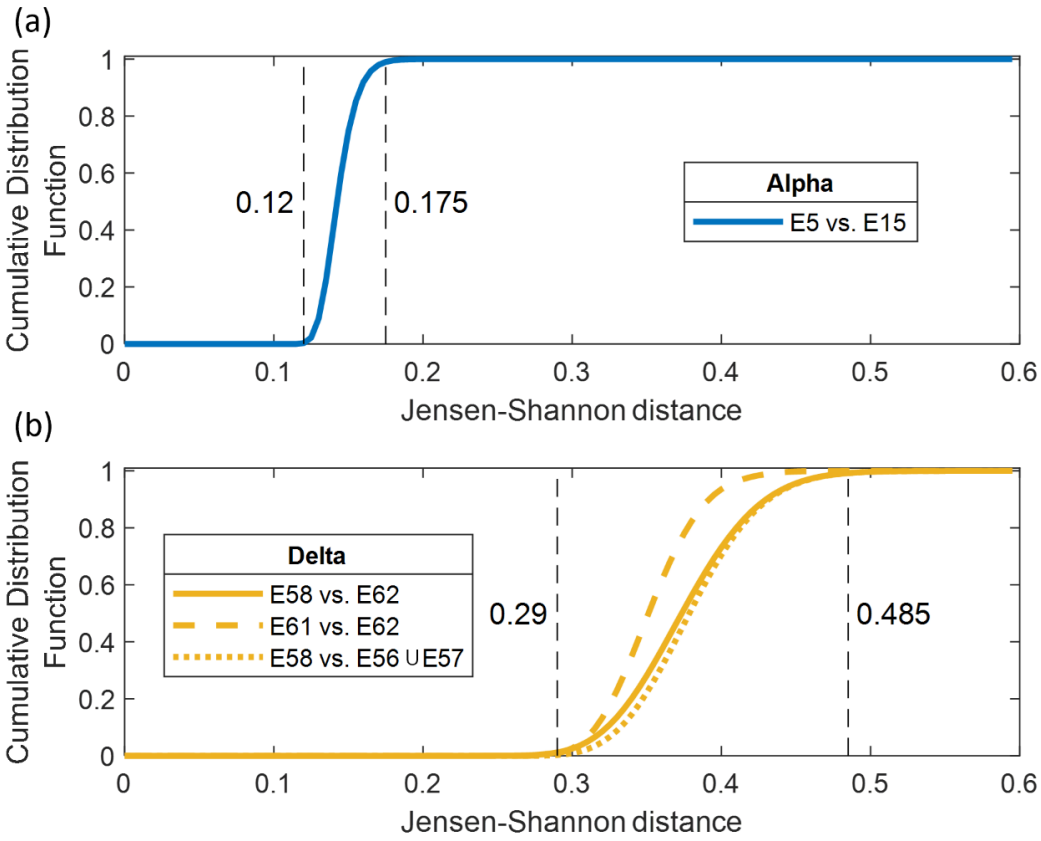


Figure 3-13. Cumulative distribution functions (CDF) of Jensen-Shannon distances for (a) Alpha variant datasets E5 and E15, both tested at 0.20 V and (b) Delta variant datasets E58 and E61, both tested at 0.20 V. For the Alpha variant, since E5 is most confused with E15 (see Figure 3-10), we show the CDF of the Jensen-Shannon distance between E5 and E15. For the Delta variant, since E58 is very occasionally confused with (E56 ∪ E57) and E62 (see Figure 3-12), we show the CDF of the Jensen-Shannon distance between E58 and (E56 ∪ E57), and, between E58 and E62. Additionally, since E61 is very occasionally confused with E62 (see Figure 3-12), we show the CDF of the Jensen-Shannon distance between E61 and E62.

3.3 Summary

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 chapter, 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 in a wide range of molecular conductance data [127].

Experimentally, the distance between the metal electrodes is varied during an SMBJ run, and conductance is measured as a function of time. As a result of the intrinsically noisy nature of the experiments, the 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 and without averaging the conductance distribution over experimental parameters. In conjunction with 2D distributions, we have experimented with a convolutional neural network paired with an XGBoost classifier. Our investigations reveal that: (i) averaged conductance distributions have a significant impact on classifier accuracy, boosting it by as much as 21%, (ii) 2D conductance distributions are beneficial for some sequences, but not all, and (iii) the backend XGBoost classifier whose input is data from a convolutional neural network provides better accuracy than a densely connected classifier 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 prima facie 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 thesis should be of use in analyzing arbitrary single-molecule conductance data for sequence identification.

4 Conductance Segments Extraction with Piecewise Linear Approximation Method

In this paper, we propose an approach based on combining the Piecewise Linear Approximation (PLA) method, machine learning models (ensemble learning method XGBoost), and neural networks (CNN as feature extractor). Different from the preprocessing procedures that were previously developed by our group, the PLA could remove noise content more effectively and keep all useful content unchanged. The resulting conductance histograms, which are constructed purely from plateau segments, provide up to 4.22% improvement in average classification accuracy. Then XGBoost and CNN methods could focus on learning essential conductance features instead of some random features created by noise, which output an average accuracy of 96.31% compared with 83.96% of the previous ML classification system. Especially with a smaller sample size, as few as 5, our classification system is able to maintain its high accuracy by over 80% for most variants (excluding Beta_MM1 and Beta_MM3 variants). Impressively, using a constant 30 sample size, classification accuracy of the exceptional variants could increase from about 83% to 97% by tuning cutoff slope, which is a parameter introduced by the PLA method. The approach proposed in this paper could significantly improve the efficiency and accuracy of SMBJ methods, making them more viable for practical applications.

4.1 Methodology

Broadly speaking, an SMBJ experiment involves capturing a molecule between two electrodes and measuring the current passing through the molecule in response to an applied voltage bias. Figure 4-1 shows an idealized setup of SMBJ experiments: A large number of identical molecules are placed on a gold substrate (plane electrode) and a gold tip (atomically sharp electrode) is pulling up repeatedly with thousands of iterations. By applying a constant voltage bias, the system will measure the current passing between the gold substrate and the gold tip during each iteration. Within one iteration, the measurement starts with the gold tip touching the gold substrate, which reads a large saturation current. Then, the gold tip pulls up at a constant velocity until the measured current is below the lower limit of the preamplifier. If no molecule is captured during one iteration, the recorded current would drop exponentially (see Figure 4-1). These decay segments contain most likely noises, instead of real conductance values obtained from target molecules. On the other hand, the captured molecule would form an electrode–molecule–electrode structure, which measures the electronic characteristics of the molecule. Theoretically, these characteristics are not affected by small displacement changes (i.e. gold tip pulling). Therefore, the current passing through the molecule will stay the same (plus some noise) across a period of time. We call these relative constant current measurements, plateau segments. Among all raw current traces, these plateau segments contain the majority of essential information (conductance values) about the target molecule.

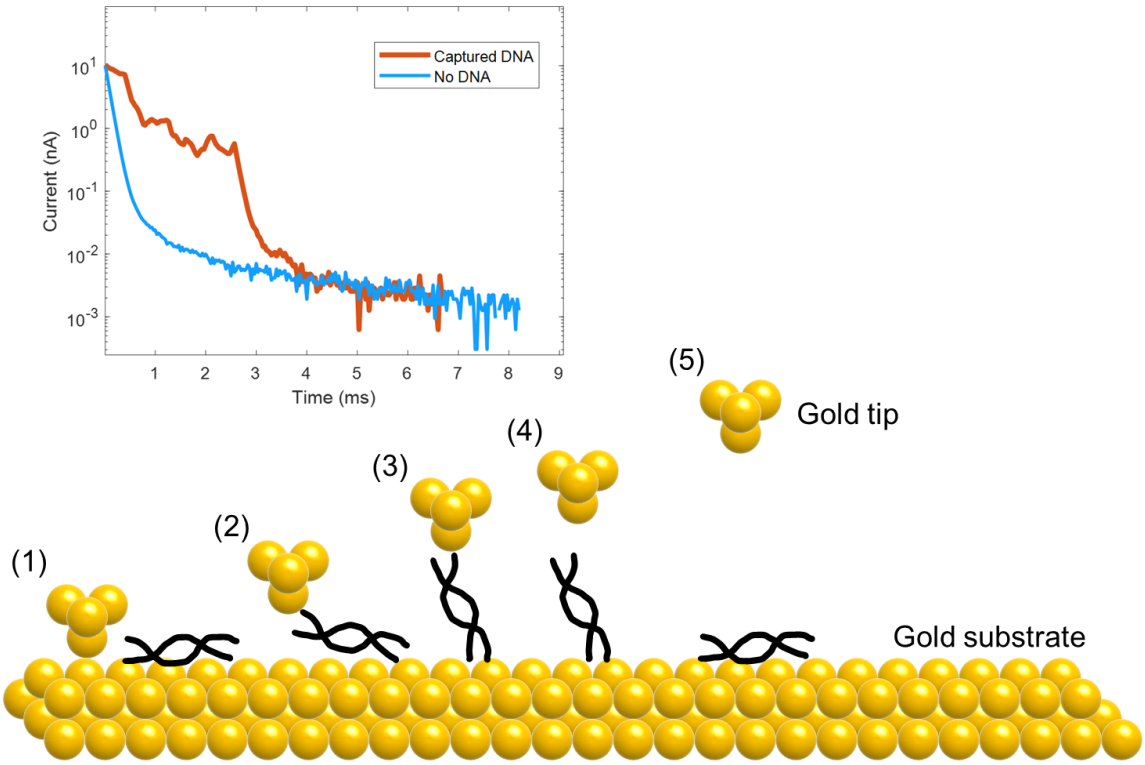


Figure 4-1. One iteration of the SMBJ measurement. A trace that captures nothing will directly go from step (1) to step (5). A trace that captures a DNA molecule, will go through all (1) – (5) steps. The plateau segment is most likely to occur between steps (2) and (3).

As discussed in the Introduction, plateau segments act as direct features of molecular interaction, free from the decay seen when no molecule is bound. Efficient identification and classification of these plateau segments is essential to improve the signal-to-noise ratio and reduce the number of measurements needed, thus making SMBJ methods more practical for real-world applications. We adopted the Optimal Piecewise-Linear Function Approximation [97] as the most suitable method for extracting plateau segments from SMBJ traces. This method uses piecewise linear functions to approximate nonlinear functions for which only sample points are available. In the section below, we will give a brief overview of this method. For further details, please refer to [97].

In the context of SMBJ measurements, a raw conductance trace is some discrete nonlinear function of time. However, as shown in Section 3.1.2, it can be remapped to an equivalent function $f(d) = g$, where g denotes conductance and d denotes the tip distance from the substrate (discretized, $d \in \{d_1, d_2, \dots, d_n\}$, where n denotes the number of samples of g). Throughout this section, we will work with conductance as a function of tip distance representation. Since the conductance values span several orders of magnitude, it is customary to work with log conductance values, i.e., the actual conductance values are 10^g. Let $\mathbb{P}$ denote the ordered set of n samples of a single experimental conductance trace, i.e., $\mathbb{P} =$

$\{(d_1, g_1),\ (d_2, g_2),\ \ \ldots, (d_n, g_n)\}$, where $d_1 < d_2 < \ldots < d_n$. The objective of the linear approximation/segmentation method is to find a set of S linear segments, $\mathbb{L} \in \{L_1, L_2, \ldots, L_S\}$, that best approximates $f(d) = g$. Each linear segment L_s is used to approximate the corresponding subset of m number of samples from $\mathbb{P}$, what we will denote by $\mathbb{P}_s$. Since the lengths of the linear segments are not necessarily identical, we will use a subscript s on m, i.e., m_s denotes the number of time instants in the segment L_s, or alternately, in $\mathbb{P}_s$. To summarize:

$$\begin{gathered}\mathbb{P}_s = \left\{\left(d_{k(s)}, g_{k(s)}\right),\ \left(d_{k(s)+1},\ g_{k(s)+1}\right),\ \ldots\ ,\left(d_{k(s)+m(s)-1},\ g_{k(s)+m(s)-1}\right)\right\} \\ L_s = \left\{(x, y):\ y = a_s x + b_s,\ \ x \in \left\{d_{k(s)},\ d_{k(s)+1},\ \ldots\ d_{k(s)+m(s)-1}\right\}\right\}\end{gathered} \tag{4-1}$$

where $k(s)$ and $k(s) + m(s) - 1$ are the distance indices of the first and last points in $\mathbb{P}_s$, (a_s, b_s) are the slope and intercepts of the linear segment L_s, and $s = 1, 2, \ldots, S$. We will require that successive segments "overlap in distance", i.e., the last point of the s^{th} linear segment is identical to the first point of the $(s+1)^{th}$ segment, $d_{k(s)+m(s)-1} = d_{k(s+1)}$. However, we do not require that our piecewise linear approximation (PLA) method be continuous at the breakpoints.

There is a limited number of molecular bonding (plateaus on a conductance trace) that can occur during one SMBJ experimental run. A careful examination of our data reveals that any conductance trace has a maximum of three plateaus and four relatively steep conductance transitions, two of which are between the three plateaus, and the other two account for possible initial (i.e., before the first plateau) and final (i.e., after the last plateau) transitions which we have observed in our data, leading to a total of seven linear segments. In our algorithm, we will therefore use $S \leq 10$. To find the best piecewise-linear approximation for a specific value of S (for $S = 1$, this is the linear least squares problem, for $S > 1$, it is known as the segmented linear least squares problem), we minimize the residual or sum of squared error (SSE), as shown below:

$$\min E = \sum_{s=1}^{S} \sum_{j=k(s)}^{k(s)+m(s)-1} \left(g_j - a_s d_j - b_s\right)^2 \ \text{w.r.t}\ \{a_s, b_s: 1 \leq s \leq S\} \tag{4-2}$$

Since the first point of the $(s+1)^{th}$ linear segment is identical to the last point of the s^{th} segment, eq. (4-2) can rewritten as follows:

$$\min E = \sum_{s=1}^{S} \sum_{j=k(s)}^{k(s+1)} \left(g_j - a_s d_j - b_s\right)^2 \ \text{w.r.t}\ \{a_s, b_s: 1 \leq s \leq S\} \tag{4-3}$$

The optimum SSE is a monotonically decreasing function of S, i.e., SSE $\to 0$ as $S \to n - 1$. In order to strike a balance between the residual error and model parsimony (the number of segments), a penalty term of the form λS, where λ is some tunable non-negative multiplier and S is the number of linear segments, can be added to the cost function in eq. (4-3). For instance, a penalty term of the form λS, where λ is some tunable non-negative multiplier and S is the

number of linear segments. However, we do not include such a penalty term and instead adopt an early termination criterion which is explained subsequently.

For $S = 1$, it is well known that the solution of eq. (4-3) is given by:

$$a_1 = \frac{\overline{d^T g} - \bar{d}\bar{g}}{\overline{d^2} - \left(\bar{d}\right)^2}, \quad b_1 = \bar{g} - a_1 \bar{d} \tag{4-4}$$

where $\bar{d} = \left(\frac{1}{n}\right)\sum_{j=1}^{n} d_j$ and $\bar{g} = \left(\frac{1}{n}\right)\sum_{j=1}^{n} g_j$ are the sample means of the tip distance values $\{d_1, d_2, \ldots d_n\}$ and conductance values $\{g_1, g_2, \ldots g_n\}$ respectively, $\overline{d^T g} = \left(\frac{1}{n}\right)\sum_{j=1}^{n} d_j\, g_j$, and $\overline{d^2} = \left(\frac{1}{n}\right)\sum_{j=1}^{n} d_j^2$ is the sample second moment of distances. For $S \geq 2$, the minimization problem in eq. (4-3) can be solved using the following dynamic programming recursion:

$$F(j, S) = \min_{S \leq i \leq j} \left(F(i, S-1) + E(i, j)\right), \quad \text{(see Figure 4-2)} \tag{4-5}$$

where $F(j, S)$ is the minimum cost to approximate the set of j points $\{(d_1, g_1),\ (d_2, g_2), \ldots (d_i, g_i),\ \ldots,\ (d_j, g_j)\}$ with exactly S segments and $E(i, j) = \sum_{k=i}^{j} (g_k - a x_k - b)^2$ is the minimum SSE incurred when the points $\{(d_i, g_i),\ (d_{i+1}, g_{i+1}), \ldots, (d_j, g_j)\}$ are approximated by a single linear segment. Assuming that the cost of approximating any set of points by 0 segments is 0, the minimum cost of approximating at least two points by one segment is $F(j, 1) = E(1, j)$, for all $j > 1$. A pseudocode and the proof of correctness of the recursion shown in eq. (4-5), which runs in $O(n^3)$ time (assuming $n > S$), is available in [97].

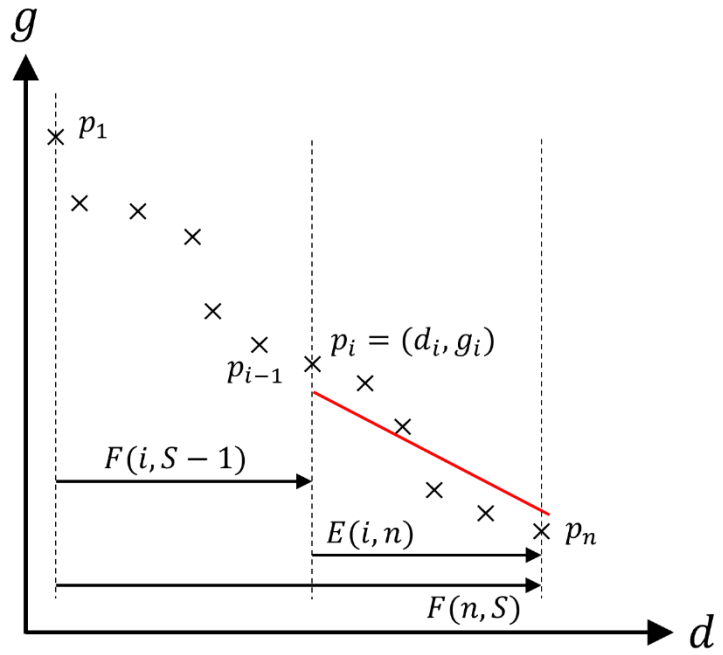


Figure 4-2. Suppose the red line in the figure is the last segment of the optimum S-segment PLA of the n points shown by a '×'. Then, the cost of the optimum PLA is given by $F(n,S) = F(i,S-1) + E(i,n)$, where $F(i,S-1)$ denotes the optimum $(S-1)$-segment PLA of the points $\{p_1, p_2, \ldots p_i\}$ and $E(i,n)$ denotes the residual corresponding to the best linear approximation of the points $\{p_i, p_{i+1}, \ldots p_n\}$. In general, the best S-segment PLA of the j points $\{p_1, p_2, \ldots, p_i, p_{i+1}, \ldots p_j\}$ is given by $F(j,S) = \min_{S \le i \le j}(F(i,S-1) + E(i,j))$.

As indicated previously, the usual way of controlling the model parsimony is to introduce a penalty parameter in the cost function which is proportional to the number of segments. However, we start the recursive process with $S = 1$ and adopt an early termination criterion,

$$\frac{F(n,\ s-1) - F(n,s)}{F(n,s)} \le \beta \tag{4-6}$$

where β is some user-defined tolerance parameter. Lower values of β result in a larger number of recursive steps, which translates to piecewise linear approximations with a larger number of segments, and *vice versa*. Based on extensive experimentation, we have found that, for our dataset, $\beta = 0.7$ provides a reasonable balance between the goodness of fit and model parsimony. Figure 4-3 illustrates the PLA method for different values of β, and corresponding number of liner segments S.

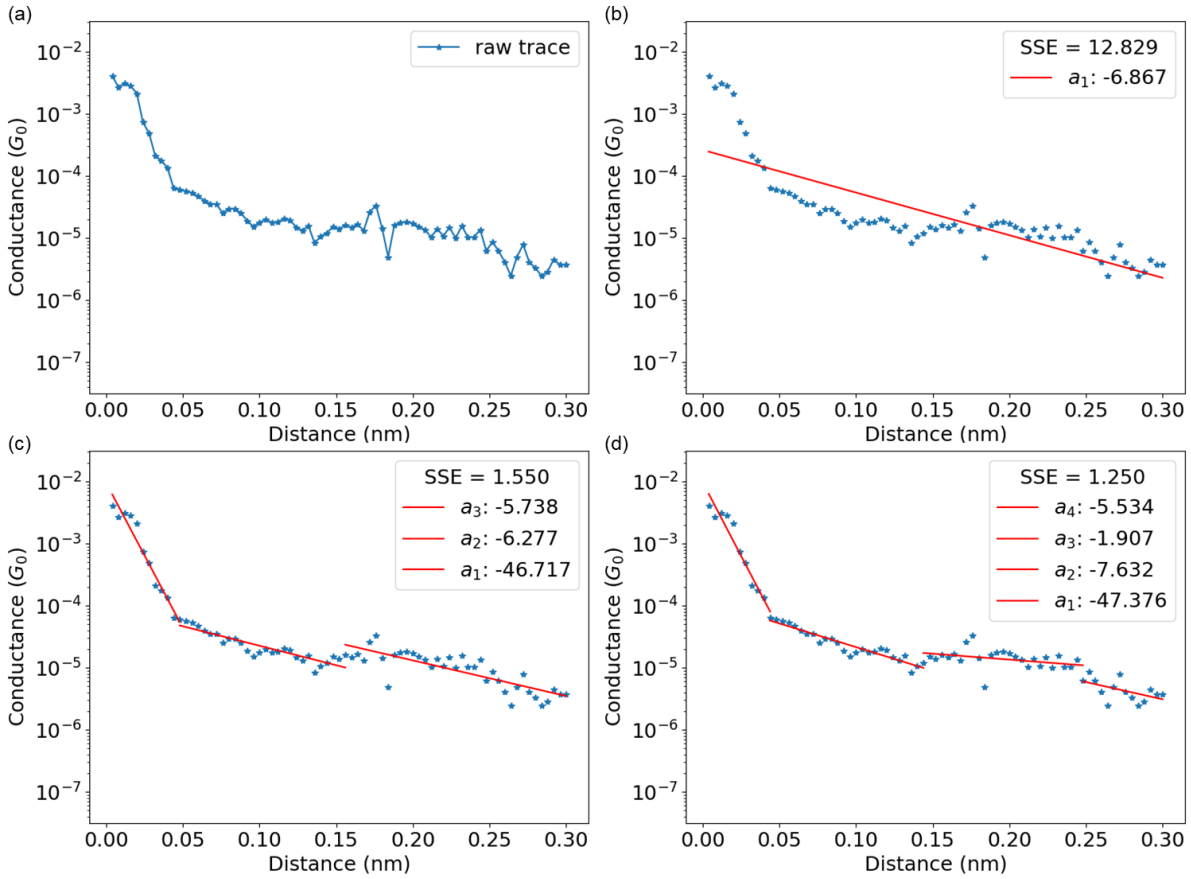


Figure 4-3. (a) Representative raw conductance trace. For panel (b), (c), and (d), the slopes of the segments and the SSE's are indicated in the panel legends. For instance, a_1 is the slope of the leftmost segment, segment numbers increasing from left to right. (b) With $\beta = 6.0$, PLA method outputs one segment for the representative raw conductance trace. (c) With $\beta = 0.7$, PLA method outputs three segments. (d) $\beta = 0.3$, PLA method outputs four segments.

Recall that our hypothesis is that the primary "information-bearing" segments of a conductance trace are its plateaus or plateau-like regions, while the steep transitions primarily contribute to noise when conductance histograms are constructed from a batch of traces. Logically therefore, after a conductance trace has been approximated by a piecewise linear curve, we elect to keep only those portions of the trace that correspond to linear approximations with absolute values of slopes smaller than a prescribed threshold. We call this threshold the "cutoff slope", denoted by A_{co}. Referring to eq. (4-1), the set of points $\mathbb{P}_s$ in the conductance trace is retained (for histogram computation) only if $|a_s| < A_{co}$. We refer to segments in the conductance trace that meet this criterion as *eligible segments*. Suppose $A_{co} = 6$. Referring to Figure 4-3, if the recursion process terminates with one linear segment (panel b), we see that the entire conductance trace is eliminated (i.e., this trace will not be used for histogram computation). However, if the recursion process terminates with three linear segments (panel c), only the portion of the trace that corresponds to the third linear segment with absolute value of slope $|a_3| = 5.738$ is eligible (i.e., only the portion of the trace corresponding to the third linear segment will be used for histogram computation). Similarly, if the recursion process terminates with four linear segments (panel d), only the portions of the trace that correspond to the third and fourth linear segments are eligible.

As discussed in the previous paragraph, for a given value of the cutoff slope, entire traces can be deemed ineligible, or certain segments of a trace can be deemed ineligible. A trace is retained if it contains at least one segment that meets the cutoff slope criterion; otherwise, it is eliminated. Lower values of the cutoff slope decrease the probability of a trace being retained (or conversely, increase the probability of a trace being eliminated). Figure 4-4 (a) shows the percentage of all Beta variant conductance traces in our database which are retained as a function of the cutoff slope. For example, when $A_{co} = 4$, we see that about 50% of all traces for the mismatch sequences have at least one eligible segment and are therefore retained (in contrast, about 90% of all traces for the perfect match sequence are retained). As $A_{co} \to 100$, almost all traces for all Beta sequences have at least one eligible segment and are retained. Figure 4-4 (b, c) shows the percentage of the number of eligible segments per retained conductance trace. For instance, when $A_{co} = 4$, about 80% of the conductance traces retained have a single eligible segment, about 18% have two eligible segments, and about 2% have three or more eligible segments. While Figure 4-4 is specifically for the Beta sequences, similar observations hold for the Alpha and Delta sequences. Since approximately 98% of the eligible traces for all sequences have either one or two plateau-like segments, we chose $A_{co} = 4$ for all baseline simulations discussed in Section 4.2.

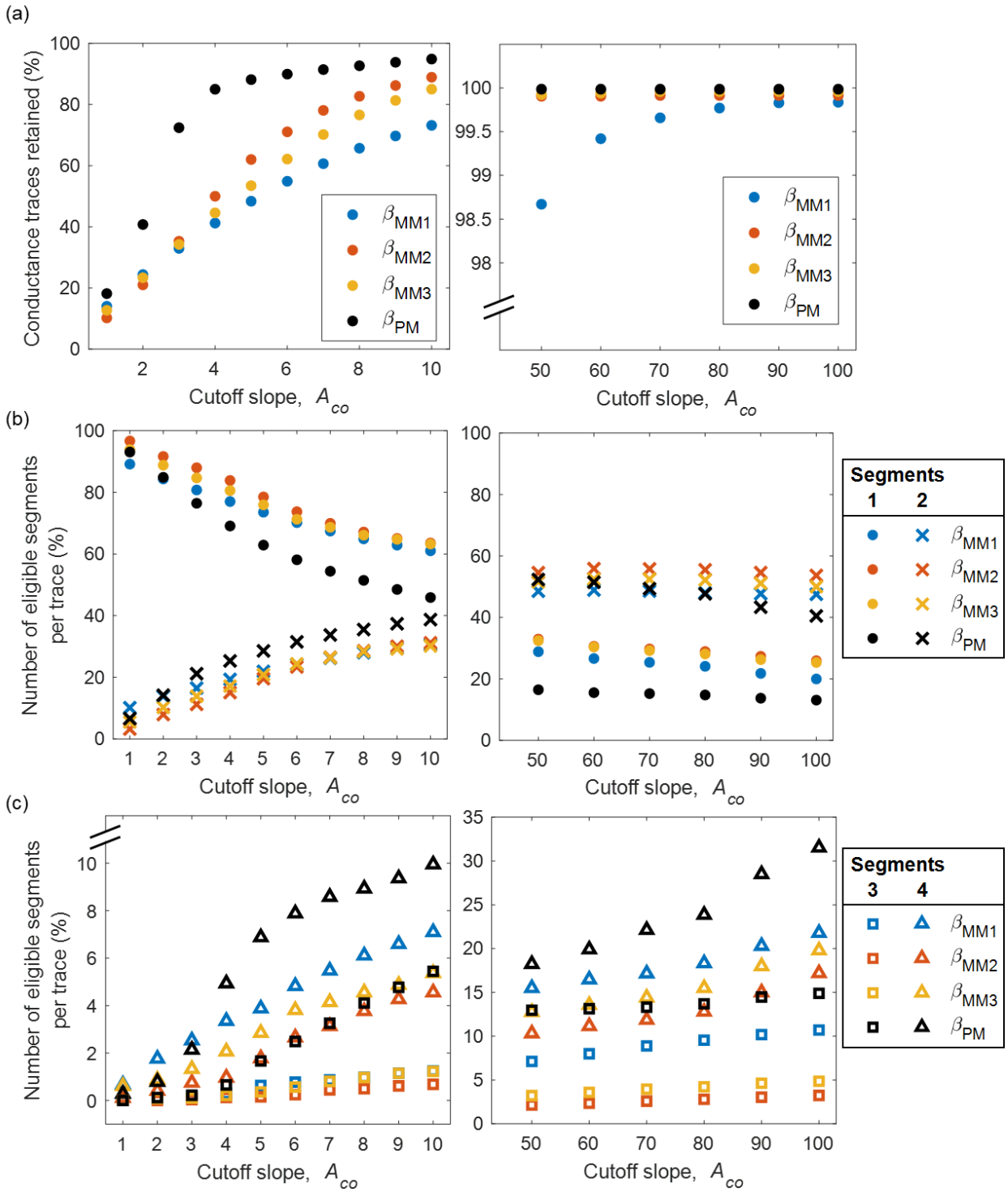


Figure 4-4. The effect of cutoff slope on: (a) the percentage of conductance traces retained and (b) the percentage of number of eligible segments per retained conductance trace. A segment is deemed eligible if the absolute value of its slope is smaller than the cutoff slope.

Figure 4-5 illustrates our overall workflow. In contrast to our previous work [38], [129], and Figure 3-3 which used an R^2 test to eliminate an "invalid conductance trace" (arising when molecule is bound to the tip, resulting in an exponential decay of the current/conductance) and a low pass filter to reduce the noise, we adopt the PLA method which plays the dual role of a selective trace/trace segment retention mechanism and a low pass filter. Notice that the filtration role is achieved since only those conductance segments whose slopes are less than the specified cutoff value are retained for histogram computation.

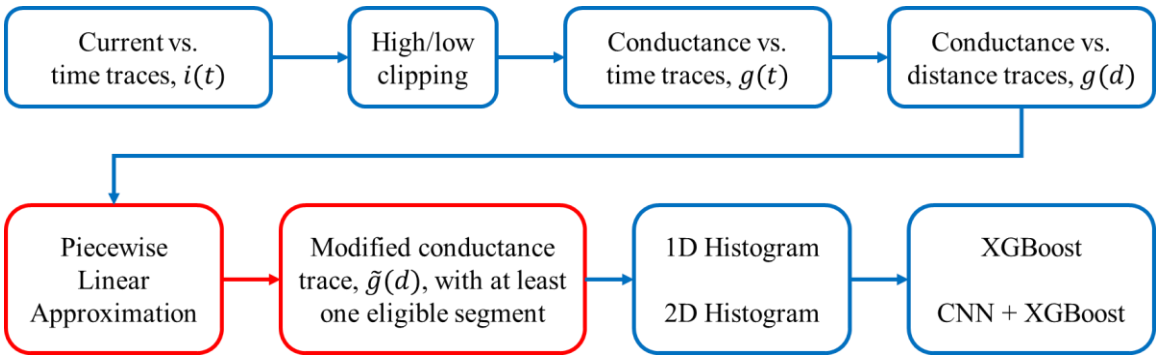


Figure 4-5. Flowchart illustrating our overall workflow. Please note that $\tilde{g}(d)$ is eliminated if there is no eligible segment after PLA preprocessing procedure.

Figure 4-6 shows the PLA method removed the noise within the edge region of conductance histograms, where steep conductance transitions commonly occurred. It also removes the $10^{-5.5}\sim10^{-4.5}G_0$ noise and shows the conductance peak more clearly in the desirable region. Additional details regarding the role of the high/low clipping, the procedure for converting a $g(t)$ trace to a $g(d)$ trace, and 1D/2D histogram computation are provided in Section 3.1.2 and 3.1.3. Appendix Figure *B-7*, Appendix Figure *B-8*, Appendix Figure *B-9* show the resulting 1D histograms, and Appendix Figure *B-10*, Appendix Figure *B-11*, Appendix Figure *B-12* show the resulting 2D histograms. In the next section, we discuss the performance of classifiers when histograms (inputs to the classifiers) are computed using the modified conductance traces obtained after the PLA preprocessing procedure.

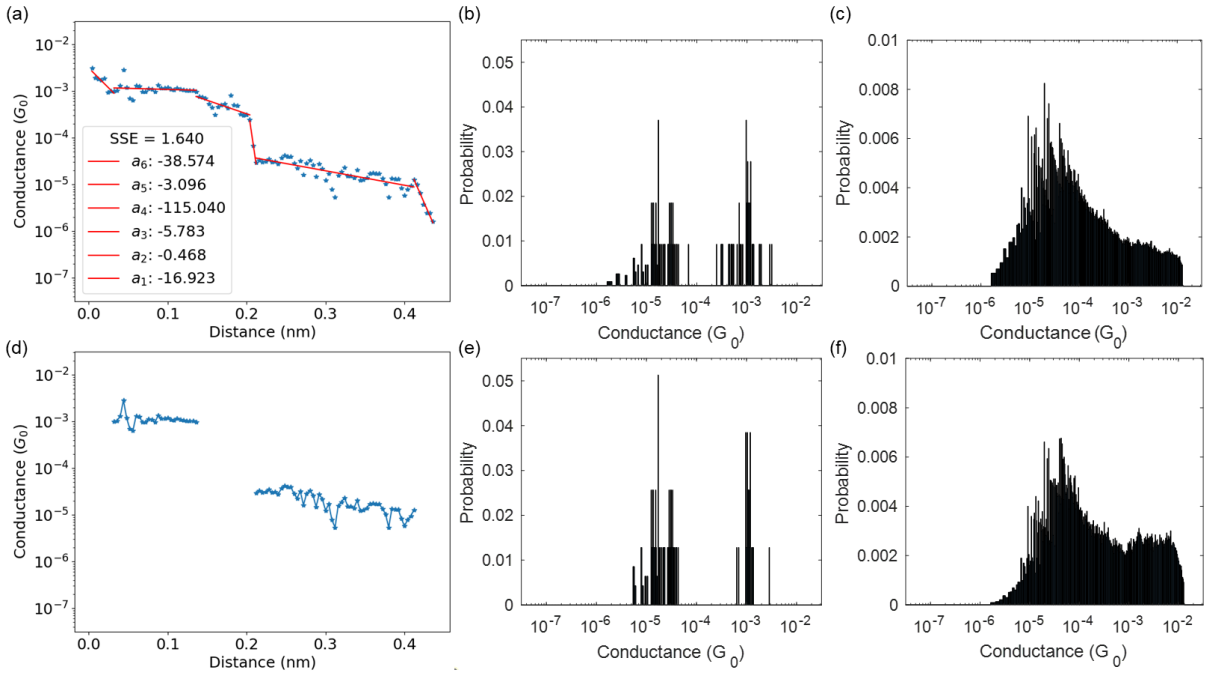


Figure 4-6. PLA procedure removes ineligible segments affecting conductance histograms. (a) PLA procedure outputs six segments for a representative raw conductance trace. (b) The (one sample) histogram of this trace. (c) The empirical large sample histogram for entire dataset of this trace. Panel (d), (e), and (f) are after removal of ineligible segments. (d) The second and fifth segments are retained with $A_{co} = 4$. (e) The (one sample) histogram of the retained trace. (f) The empirical large sample histogram for entire dataset of the retained traces (only eligible segments).

4.2 Results and Discussions

Recall from the previous section that, given a conductance trace $g(d)$, the output of passing it through the piecewise linear approximation (PLA) method is either a rejected trace or a modified conductance trace $\tilde{g}(d)$ with segments which can be approximated by a linear function with an absolute value of slope less than a prespecified threshold, A_{co}. Instead of using the set of raw conductance traces $g(d)$, the set of modified conductance traces $\tilde{g}(d)$ is used for constructing conductance histograms, which can be either 1D or 2D (explained subsequently). In this section, we will evaluate the effectiveness of the PLA method on extracting suitable conductance features from SMBJ traces, which ultimately affects the accuracy of classifier models.

To ensure consistency with our previous work, we have used the same dataset that we previously used in Section 3. The dataset contains ten unique 12 base-pair DNA segments from one of three variants of the SARS-Cov-2 genome. The three variants are Alpha B.1.1.7, Beta B.1.351, and Delta B.1.167 [101]. There are a few experiments performed for each of the ten DNA sequences, and each experiment contains several thousands of current traces. As shown in

Table 3-1, there are 39 experimental datasets in total for the ten DNA sequences, which includes a choice of experimental parameters (ramp rate, current amplifier sensitivity, and applied bias). Of the ten DNA sequences, three are perfectly matched with the target Alpha, Beta, and Delta variants, referred to as Alpha_PM, Beta_PM, and Delta_PM (perfect match), respectively. The others correspond to sequences of the same length but with a single base pair mismatch, referred to as mismatches (MM). The mismatch sequences are denoted by Alpha_MM1, Alpha_MM2, Beta_MM1, Beta_MM2, Beta_MM3, Delta_MM1, and Delta_MM2. Table 3-1 shows the exact mutation locations for these sequences. For detailed discussion about experimental rationale and source of conductance data please refer to [129] and [91].

As recorded, each conductance trace is a function of time, which can be converted in terms of the distance between the two electrodes. We divide the range of distance values, $[0, 0.4]$ nm, into a number of distance bins, which we denote by $N_{\text{bins\_Distance}}$. Similarly, the range of conductance values, $[10^{-7.5}, 10^{-1.5}]$ G_0, is divided into a number of conductance bins, which we denote by N_{bins}. We chose N_{bins} = 600 for both 1D and 2D histograms and $N_{\text{bins\_Distance}}$ = 10 for 2D histograms (equal to 1 for a 1D histogram). Additional details regarding 1D/2D histogram construction are available in Section 3.1.3.

Figure 4-7 shows the empirical large sample (constructed from a large number of conductance traces) 1D/2D histograms of a representative dataset, for three different preprocessing procedures: (i) no preprocessing (i.e., all experimentally recorded traces are used to construct the 1D/2D histograms), shown in panels (a) and (d), (ii) R^2 test + LPF preprocessing procedure adopted in our previous work [38], which is intended to filter out "invalid" traces or those with no molecular binding (i.e., all experimentally recorded traces which pass the R^2 test with a threshold of 0.95 are used to construct the 1D/2D histograms), shown in panels (b) and (e), and (iii) with PLA preprocessing procedure instead of R^2+LPF (i.e., all experimentally recorded traces which pass the PLA method with at least one eligible segment are used to construct the 1D/2D histograms), shown in panels (c) and (f). From Figure 4-7 (a) and (d), we can observe that the histogram constructed from the raw conductance traces with no preprocessing is noisy and shows relatively random and ill-defined conductance features (large peaks for 1D histogram, yellow blocks for 2D histograms). In contrast, we observe from Figure 4-7 (b) and (e) that the histograms computed after R^2+LPF procedure show a bimodal distribution with peaks around $10^{-5.5}$ G_0 and 10^{-4} G_0. The PLA processed histogram shown in Figure 4-7 (c), however, has a pronounced dominant peak at approximately $10^{-5.5}$ G_0, which occurs at a distance of 0.3-0.4 nm, as can be inferred by comparing the corresponding 2D histograms shown in Figure 4-7 (e) and (f). We can therefore also infer that the conductance peak at 10^{-4} G_0 observed in Figure 4-7 (b) occurs when the tip distance is ≈ 0.1 nm, which is less likely to represent the conductance value of a 12 base-pair DNA sequence. Based on the dimension of DNA, 0.1 nm tip distance is too short for the occurrence of a DNA molecule.

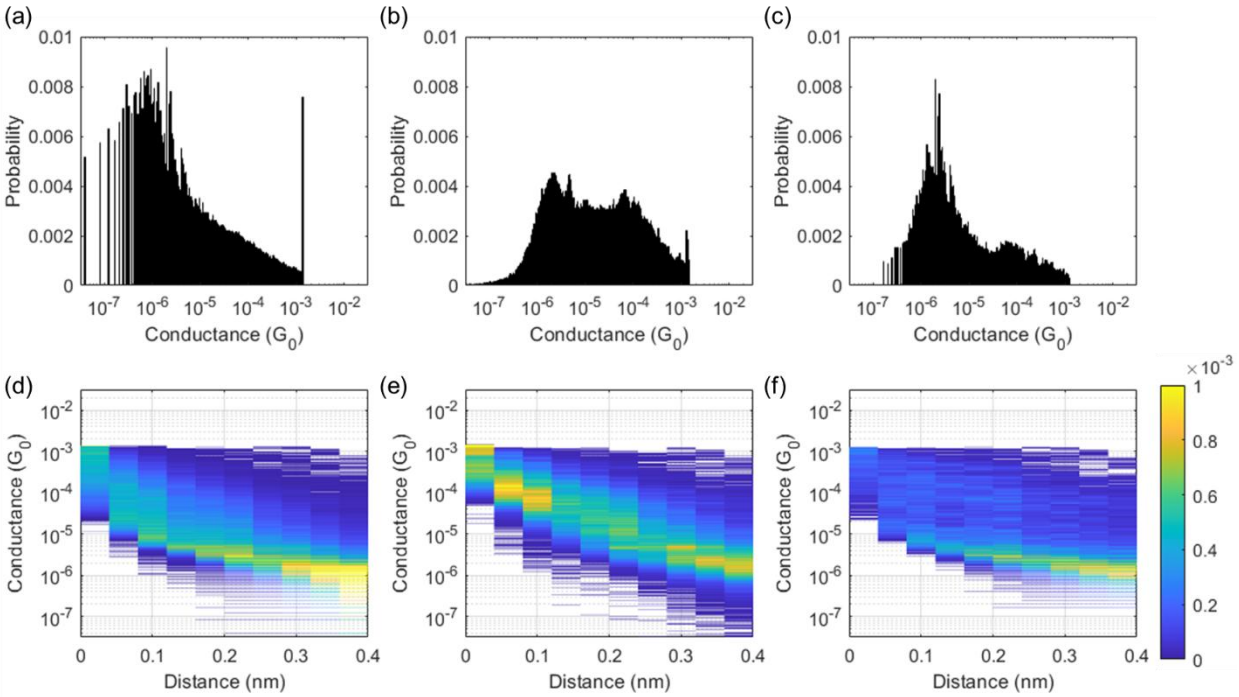


Figure 4-7. Representative empirical *large sample* histograms for the Delta_MM2 sequence, dataset E57. (a, d) 1D and 2D histograms constructed from raw conductance traces without any preprocessing. (b, e) 1D and 2D histograms with R^2 test (threshold of 0.95) + LPF procedure. (c, f) 1D and 2D histograms with PLA procedure ($A_{co} = 4$) instead of R^2+LPF.

4.2.1 Performance of baseline classifiers

Given a set of modified conductance traces $\tilde{g}(d)$ for each of the ten sequences, we sample H traces randomly and construct 1000 histograms for each sequence, 70% of which are used for classifier training and validation, while the rest are used for testing. Similar to our previous work [129], Section 3.2.1, we adopt 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. Consider the Alpha_MM1 sequence which has 5 different labels: E1, E2, E3, E4, and E5 (see Table 3-1). These labels correspond to different choices of experimental parameters, voltage bias, current amplifier sensitivity, and ramp rate. When constructing a histogram for Alpha_MM1 from H traces, we can either sample all H traces from a particular combination of experimental parameters (e.g., E3, corresponding to a bias of 0.1V, a current amplifier sensitivity of 10 nA/V, and a ramp rate of 10 V/s) or sample H traces randomly from all combinations of experimental parameters. 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*.

For 1D histograms, we have used the XGBoost classifier [120], [121], [122]. Two critical hyperparameters within an XGboost framework are the number of trees/estimators, N_{est}, and the depth of each tree/estimator, D_{est}. Based on extensive hyperparameter optimization, we chose N_{est} = 200 and D_{est} = 2. For 2D histograms, we use a Convolutional Neural Network (CNN) for feature extraction, which is paired with an XGBoost classifier (see Figure 4-13).

In this section, we discuss the performance of a *baseline classifier* which is characterized by the following parameters: (i) prefiltering with $R^2 \leq 0.95$ plus LPF, or, using the PLA procedure with $|A_{co}| = 4$, (ii) histograms from H = 30 traces, and (iii) number of conductance bins N_{bins} = 600, number of distance bins $N_{\text{bins\_Distance}}$ = 10 for 2D histograms. Table 4-1 compares the overall classifier accuracies of the two preprocessing approaches (R^2+LPF or PLA) for the four different input representations. We observe that the PLA procedure provides a substantial improvement in classification accuracy (average 2.77%) for all four input representations. Although the improvement is most pronounced for approach A3 (non-averaged 2D histograms), the best performer is approach A4, closely followed by A2, both of which use averaged histograms as input features. This consistent performance improvement underscores the effectiveness of the PLA procedure and highlights its potential as a critical step in SMBJ experimental data analysis.

Table 4-1. Summary of four different input data representations, ML algorithms used, and corresponding classification accuracies for two preprocessing approaches, R^2+LPF or PLA.

Approach	Histogram dimensionality	Average conductance histogram?	ML algorithm	Average accuracy, H = 30 (over 100 trials)	
				R^2 + LPF	PLA
A1	1D	No	XGBoost	83.96%	85.45% (+1.49%)
A2	1D	Yes	XGBoost	92.09%	95.61% (+3.52%)
A3	2D	No	CNN + XGBoost	86.40%	90.62% (+4.22%)
A4	2D	Yes	CNN + XGBoost	94.45%	96.31% (+1.86%)

4.2.2 Performance analysis of baseline classifiers w.r.t sample size, *H*

In this section, we address the impact of the sample size parameter H on the accuracy of a baseline classifier. Figure 4-8 shows the classification accuracy for the best performer from

Table 4-1, approach A4, with respect to the sample size parameter H, keeping $A_{co} = 4$, N_{bins} = 600, and $N_{\text{bins\_Distance}}$ = 10. Intuitively, we expect larger values of H to yield better classifier accuracies since the stochasticity inherent in the experimental traces will tend to be averaged out as H increases, revealing the underlying true probability distribution. From a usability perspective, however, a highly accurate classifier model operating on histograms constructed from a single trace (H = 1) is ideal since sequence determination can be made at run time after every experimental run, thereby reducing the data collection overhead. We observe that the classifier accuracies do indeed exhibit the expected trend. In fact, the accuracy is over 99% when $H = 40$ for all sequences, except the Beta_MM1 and Beta_MM3 sequences, a phenomenon that was also observed and reported in our previous work [129] which was based on R^2+LPF procedure. Other than the two Beta variants, accuracy over 90% can be achieved even with sample values as low as $H = 10$, which takes us close to our desired goal. In Section 3.2.2, we remarked "… 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." Clearly, this statement stands invalidated if the PLA procedure approach is adopted, in which case approach A4 outperforms approach A2 for every single sequence (see Figure 4-10).

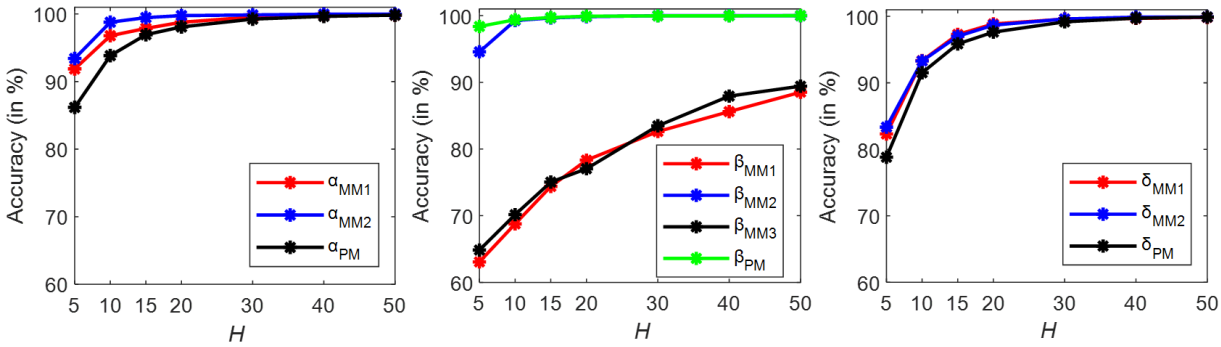


Figure 4-8. Performance analysis of baseline classifier with respect to H for approach A4.

In order to better understand the role of the two preprocessing methods, we compare the classifier accuracies for the ten sequences when $H = 30$. As shown in Figure 4-11 (the confusion matrices used to generate this figure are shown in Figure 4-9 and Figure 4-10), the PLA preprocessing procedure is *almost* universally superior to our previous R^2+LPF preprocessing approach, the most significant beneficiaries being the Delta group variants, in particular Delta_PM and Delta_MM1, followed by Delta_MM2. In Section 3.2.2, we reported that "In general, we have found that Delta variants are harder to classify than Alpha variants." As Figure 4-11 shows, this statement is no longer valid, and both the Alpha and Delta variants can be detected with a very high degree of accuracy (> 99%), simply by adopting a PLA-based approach.

Figure 4-9 shows the confusion matrices of baseline classifier results for all four approaches using R^2 test and low pass filter. Figure 4-9 (a) and (b) are the same confusion matrices from the

previous publication [129]. Figure 4-9 (c) and (d) are new confusion matrices of 4 stacked CNN and XGBoost model (see Figure 4-13).

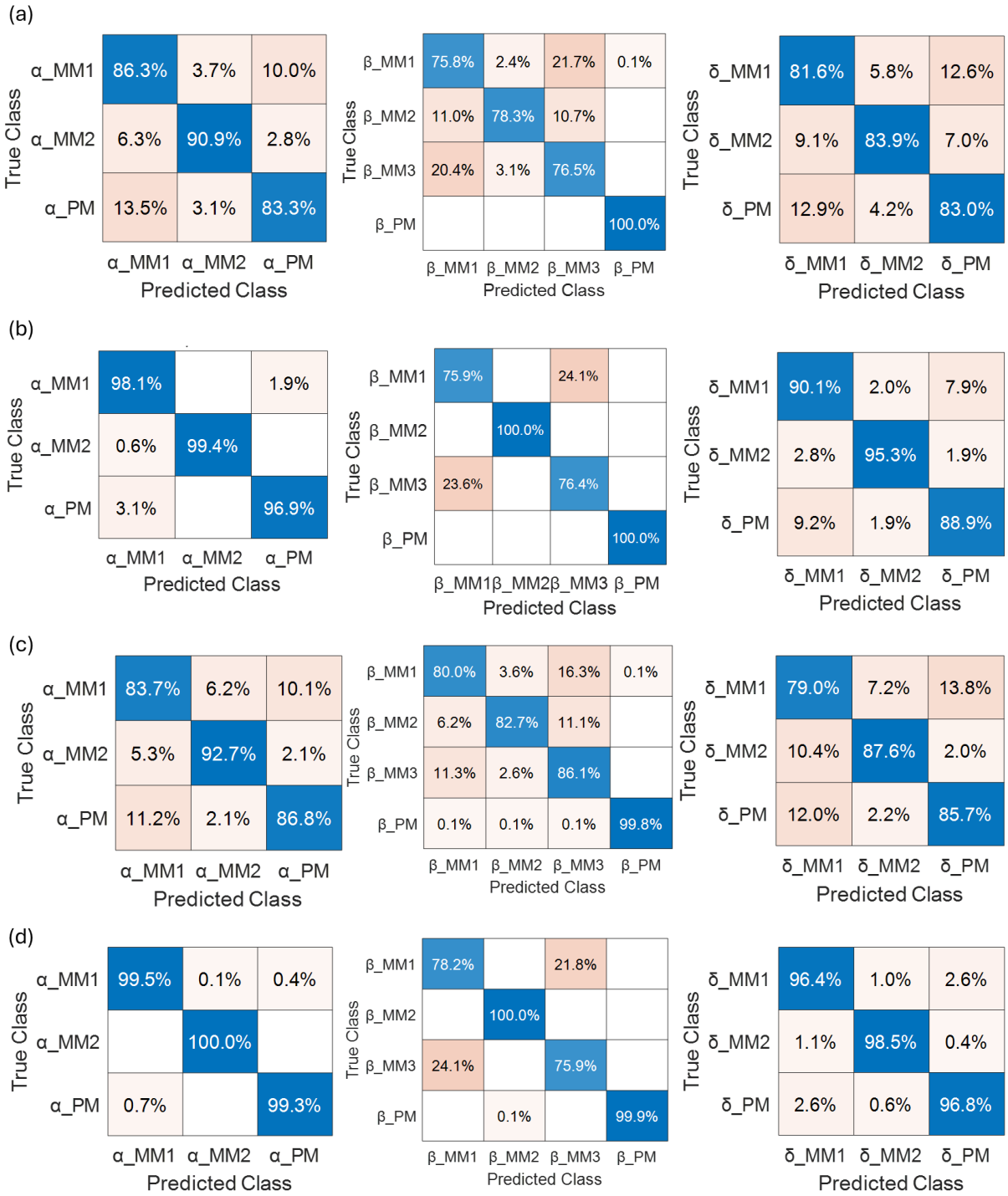


Figure 4-9. Confusion matrices for baseline classifiers: (a) Approach A1, (b) Approach A2, (c) Approach A3, and (d) Approach A4 using R^2 test and low pass filter.

Figure 4-10. Confusion matrices for baseline classifiers using Linear Piecewise Approximation method: (a) Approach A1, (b) Approach A2, (c) Approach A3, and (d) Approach A4. The differences between these four approaches are summarized in Table 4-1.

However, our previous R^2+LPF approach still turns out to be better (by 4.6%) for the Beta_MM3 sequence. Since the histograms are directly affected by the choice of the cutoff slope threshold in the PLA procedure, it is logical to ask at this point whether increasing the slope threshold, thereby retaining a greater number of linear segments, might improve the accuracy of the Beta_MM3 sequence, and to a certain extent, the Beta_MM1 sequence (notice that the PLA

procedure narrowly outperforms the R^2+LPF procedure), since these two sequences are misclassified w.r.t each other approximately 17% of the time. We will revisit this issue in the next section, after an analysis of classifier accuracies with respect to the choice of the cutoff slope.

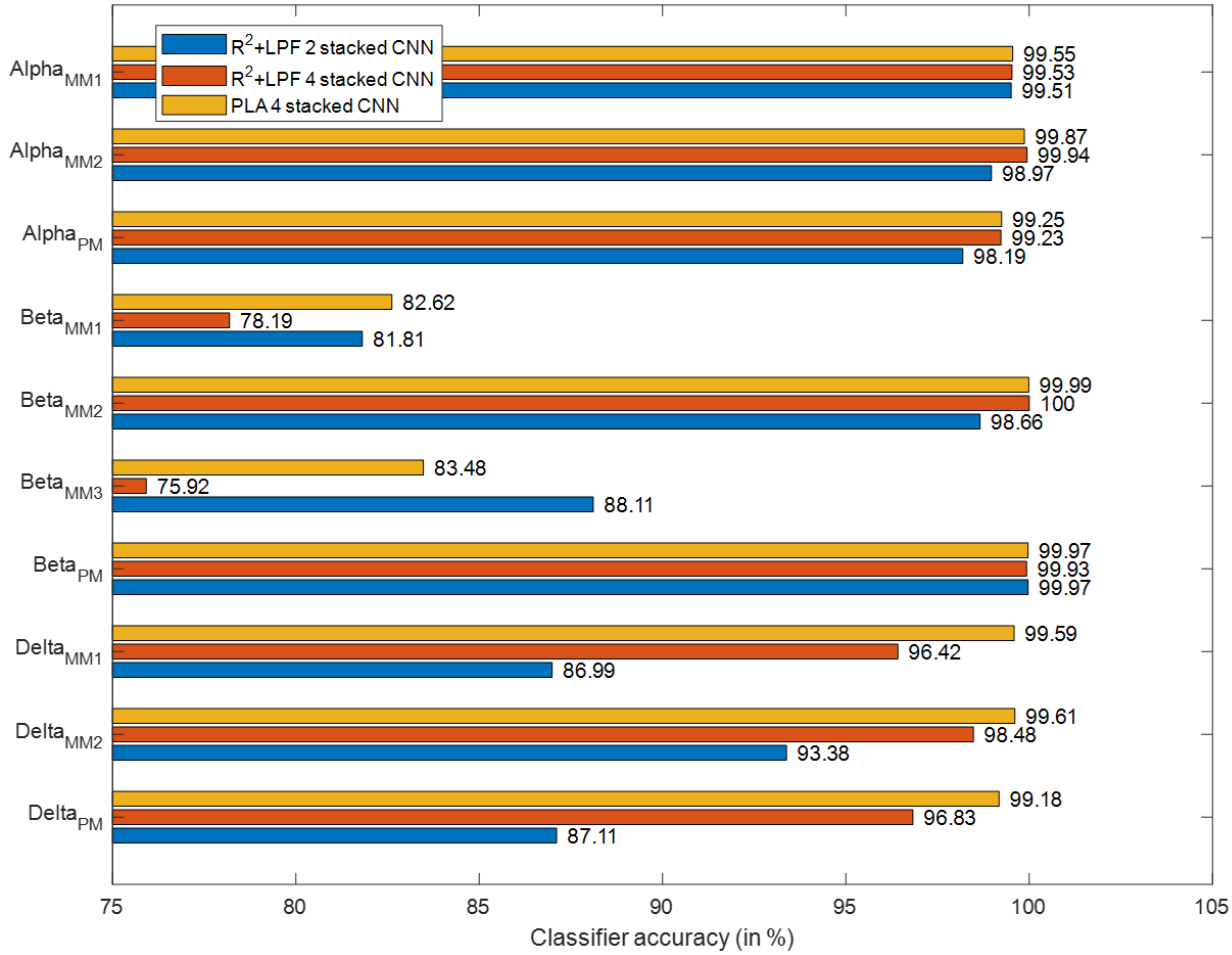


Figure 4-11. Comparison of baseline classifier accuracies for $H = 30$, approach A4, when the traces are R^2+LPF processed (previous work [129]) vs. PLA processed (current work).

4.2.3 Impact of cutoff slope on classifier accuracy

The cutoff slope threshold, A_{co}, in the piecewise linear approximation procedure is a critical parameter that directly affects the conductance histograms after the traces are processed. In Figure 4-4, we analyzed how the number of retained segments changed with the choice of A_{co}. In this section, we study its impact on the accuracy of the classifier models. Figure 4-12 shows the classifier accuracies for approach A4 as a function of A_{co}, maintaining the baseline conditions H = 30, N_{bins} = 600, and $N_{\text{bins\_Distance}}$ = 10. We observe that for all sequences, except Beta_MM1 and Beta_MM3, the optimal (w.r.t classifier accuracy) value of the cutoff slope threshold is $A_{co} = 4$, with most of the performance gain occurring between $1 \leq A_{co} \leq 3$.

Beyond $A_{co} = 4$, classifier performance saturates for most of the sequences, barring Beta_MM1 and Beta_MM3, with two sequences (Alpha_MM1 and Alpha_PM) exhibiting a slight performance dip. When the threshold value is small (e.g., 1 or 2), the linear segmentation method purges too much of the information-bearing regions of the conductance traces, regions that contribute to the accuracy of the computed histograms and therefore the classifier performance. For the Beta_MM1 and Beta_MM3 sequences, however, we observe a steady improvement in accuracy until $A_{co} = 80$. At this value of the cutoff slope, the accuracies are 96.52% and 97.58%%, which dramatically increased from 82.62% and 83.48% at $A_{co} = 4$, for the Beta_MM1 and Beta_MM3 sequences respectively. Please note that we still obtain on average 2 segments per retained traces with $A_{co} = 80$.

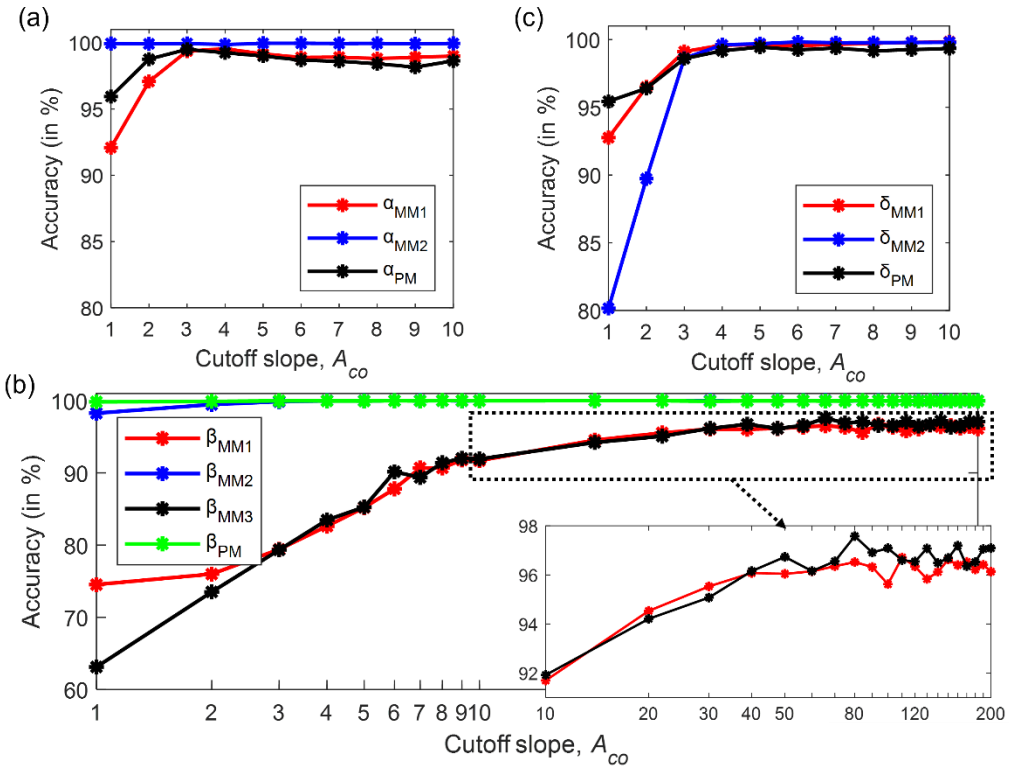


Figure 4-12. Impact of cutoff slope on classifier accuracy for approach A4. (a) For Alpha variants, classifier performance saturates after $A_{co} = 4$. (b) For Beta_MM2 and Beta_PM, classifier performance saturates after $A_{co} = 4$. For Beta_MM1 and Beta_MM3, classifier performance saturates after $A_{co} = 80$. (c) For Delta variants, classifier performance saturates after $A_{co} = 4$.

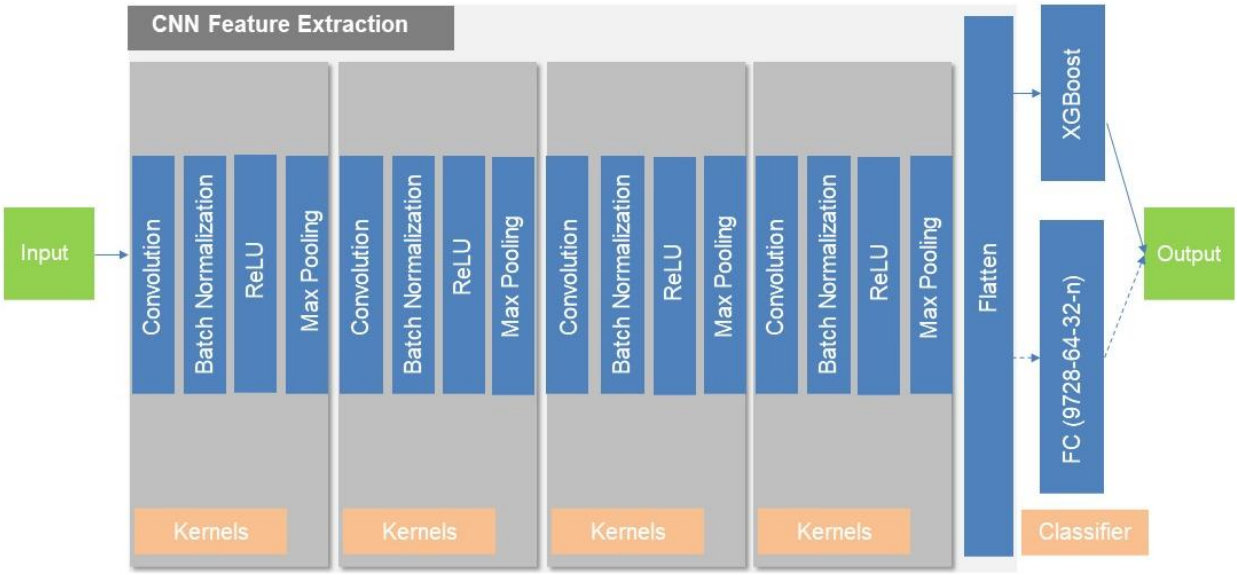


Figure 4-13. Sequence of components in the stacked CNN and XGBoost model. FC stands for fully connected layer.

We utilized the Tensorflow implementation of CNN with a learning rate based on the Adam optimizer. For the multiclass classifications, we employed categorical cross entropy as the loss function. Figure 4-13 shows the structure of the CNN feature extractor and the XGBoost model. Starting from the method in our previous approach, we identified that a CNN feature extractor with greater depth obtains superior accuracy on the linear approximation treated data. We hypothesize that this is due to the decreased noise and stochasticity in the filtered Linear Approximation data vis-à-vis the R95 filtered data, which reduces the effect of overfitting on more complex models. This allows a more complex CNN model to identify more features of the experimental data without suffering from over-generalization (see Figure 4-13). Accordingly, we added two additional sets of convolutional, batch normalization, ReLU, and Max Pooling layers to the CNN feature extractor. Finally, we passed the output of the flattening layer as the input for the XGBoost model (instead of the traditional fully connected layers) for classification into target classes.

4.3 Summary

This chapter presents a statistical and machine learning method for enhancing the accuracy and efficiency of conductance measurements in SMBJ experiments. By utilizing the PLA method, conductance plateau segments are effectively and precisely isolated, enhancing the clarity of conductance data and mitigating the noise inherent in SMBJ measurements. The resulting plateau-only conductance histograms enabled more effective training of machine learning models, resulting in substantial improvements in classification accuracy across various configurations. The combination of PLA with XGBoost and CNN-based models demonstrated an

average accuracy improvement of up to 4.22% compared to traditional low-pass filtering approaches.

These findings highlight the potential of PLA as an effective processing tool to enhance the practicality of SMBJ experiments. By focusing solely on stable conductance segments, our approach ensures that the derived features are more representative of the intrinsic molecular properties, which is crucial for identifying subtle variations between different molecular variants and conformations. Additionally, the robustness of this method across multiple machine learning configurations indicates its versatility, suggesting that PLA could be employed in a broader range of single-molecule analysis techniques, including other types of experimental setups involving molecular junctions or time-series data.

Future work will focus on optimizing the PLA for various molecular targets and expanding its application to different types of single-molecule systems, such as protein interactions, chemical sensing, and real-time biological processes. By continuing to enhance these methods, we aim to push the boundaries of what is achievable with SMBJ technology, ultimately paving the way for more practical and scalable applications in molecular electronics, nanotechnology, and biochemical sensing.

5 Role of Counterions and Solvent Dielectric on the Conductance of B-DNA

DNA naturally exists in a solvent environment, comprising of water and salt molecules such as sodium, potassium, magnesium, etc. Along with the sequence, the solvent conditions become a vital factor determining DNA structure and thus its conductance. Over the last two decades, researchers have measured DNA conductivity both in hydrated and almost dry (dehydrated) conditions. However, due to experimental limitations (the precise control of the environment), it is very difficult to analyze the conductance results in terms of individual contributions to the environment. Therefore, modeling studies can help us to gain a valuable understanding of various factors playing a role in charge transport phenomena. DNA naturally has negative charges located at the phosphate groups in the backbone, which provides both the connections between the base pairs and the structural support for the double helix. Positively charged ions such as the sodium ion (Na^+), one of the most commonly used counterions, balance the negative charges at the backbone. This modeling study investigates the role of counterions both with and without the solvent (water) environment in charge transport through double-stranded DNA. Our computational experiments show that in dry DNA, the presence of counterions affects electron transmission at the lowest unoccupied molecular orbital energies. However, in solution, the counterions have a negligible role in transmission. Using the polarizable continuum model calculations, we demonstrate that the transmission is significantly higher at both the highest occupied and lowest unoccupied molecular orbital energies in a water environment as opposed to in a dry one. Moreover, calculations also show that the energy levels of neighboring bases are more closely aligned to ease electron flow in the solution.

The rich diversity in conclusions reached depends on the experimental and modeling methods and calls for a continued systematic study of the problem. In this modeling study, we focus on investigating the roles of Na^+ ions and the solvent environment in determining the intrinsic conductance through DNA. The importance of this model study is that the results help us understand the role of solvent and counterions alone in determining the transmission without convolving the structural effects. We use the textbook forms of B-DNA with the sequence of 5′-CCCGCGCCC-3′. We chose this sequence based on previously published work [15], [130]. Our results show that Na^+ ions can significantly impact the charge transport properties of the DNA strand depending on the dielectric constant of the environment. In the dehydrated condition (low dielectric constant), the addition of Na^+ ions lowers the band gap to 0.77 eV compared with water (high dielectric constant), which has a band gap of 4.03 eV. This difference is because Na^+ ions add unoccupied energy levels in the band gap of the DNA in a dehydrated condition. In contrast, for the water solvent, Na^+ counterions add unoccupied energy levels that have higher energy than the LUMO (lowest unoccupied molecular orbital), which is primarily located on the DNA. This observation can be attributed to the high dielectric constant of water, which reduces the interaction between DNA and Na^+ ions due to the charge screening effect. To demonstrate

the generalizability of our conclusions, we applied the same simulation procedure to various DNA sequences with different lengths. Please refer to Section 5.2.5 for detailed discussions.

In addition, we find that the high dielectric constant increases the electronic coupling between the molecular orbitals of the DNA and yields a smaller on-site energy separation between them. Therefore, the transmission is at least two orders of magnitude larger at the HOMO (highest occupied molecular orbital) and LUMO regions of the DNA with the water solvent. The rest of this chapter is structured as follows. Section 5.1 discusses our simulation procedure, including density functional theory (DFT) and charge transport calculations with Green's function method. In Section 5.2, we compare energy levels, transmission plots, wave function analysis, and hopping parameters of the DNA molecule in both water and dry cases. Finally, we summarize the concluding remarks in Section 5.3.

5.1 Methodology

The simulation procedure can be broken down into three steps: obtaining the atomic coordinates of the DNA molecule and counterions, performing density functional theory (DFT) calculations, and calculating the transmission using Green's function approach. (see Figure 5-1).

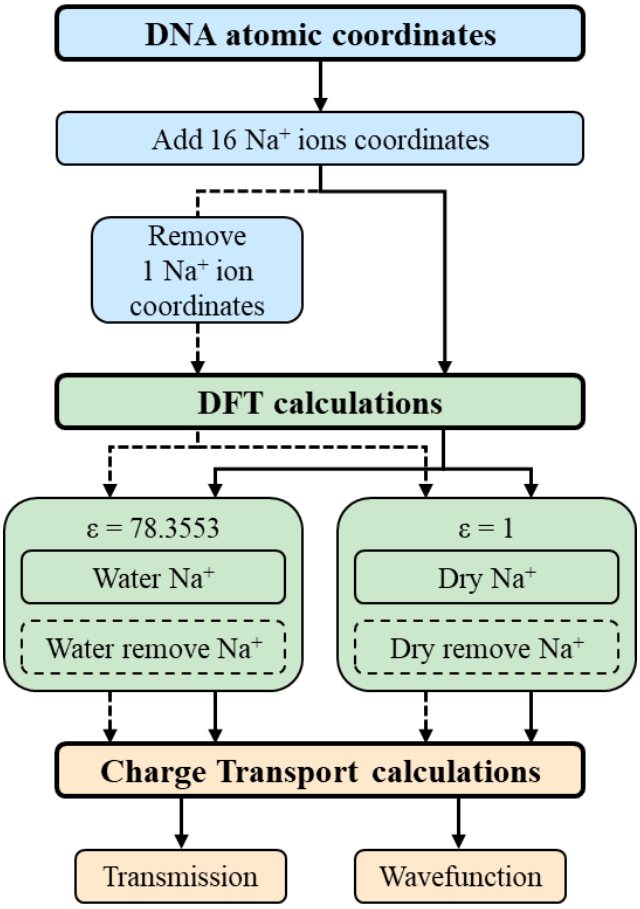


Figure 5-1. The flow chart of simulation procedures.

5.1.1 DFT calculations

We first obtain the atomic coordinates of the double-stranded B-DNA using the Nucleic Acid Builder [131]. Then, we add counterions along the backbone using the approach in Qi et al. [132]. The minimization of Na^+ ions is performed with a single strand of B-conformation DNA consisting of three bases and the phosphate backbone while the DNA is held fixed. The relative dielectric constant of unity (dry) was used. Qi et al. found that the sodium atom should be at a location of 2-3 Å away from the phosphate group (see Appendix C for the precise coordinates used). We followed this with DFT calculations to find the Fock and overlap matrices of the DNA strand with counterions. As we are probing the role of the solvent dielectric in this work, we keep the coordinates of the DNA and counterions fixed. The only difference between dry (dehydrated) and water (hydrated) DNA is the value of the dielectric constant in the DFT calculations. While energy-minimized coordinates for atoms in dry DNA would be more appropriate, force fields for the dehydrated condition are not available, and so in this computational study, we use the same coordinates in both cases. In a solvent environment, we expect the location of the counterion to be further away from the phosphate.

In calculating the impact of the solvent, *ab initio* studies have concluded that the polarizability of the solvent is the essential factor affecting ionization potential [133], [134], [135]. Studies show that the polarizable continuum model (PCM) captures the screening effect of the solvent without the need to include explicit water molecules [135]. Therefore, to account for the water solvent, we use the PCM. The DFT calculations are carried out with the B3LYP/6-31G(d,p) basis set [136] with one counterion at every phosphate backbone, with the total charge of the system being zero. We choose this basis set based on a balance between calculation accuracy and reasonable computational cost. We verified that our results are consistent with different basis sets: B3LYP/6-311G(d,p), B3LYP/cc-pVDZ, and B3LYP/cc-pVTZ. Please refer to Section 5.2.3 for detailed discussions. Note that the terminal bases at the 5′ end do not include the phosphate groups. After reaching convergence, the Fock and overlap matrices obtained from the DFT calculations are used in the transport calculations, which is the third and final step. To understand the role of counterions in influencing transport, we randomly select one of the 16 Na^+ ions and remove its coordinate right before DFT calculation. We refer to these calculations as the "removed Na^+" case, and the total charge of the system is set to -1. The simulation procedure is kept identical for all cases. Thus, overall, we investigate four different cases: *water* Na^+ (ε = 78.3553, with 16 Na^+ ions), *water removed* Na^+ (ε = 78.3553, with 15 Na^+ ions), *dry* Na^+ (ε = 1, with 16 Na^+ ions), and *dry removed* Na^+ (ε = 1, with 15 Na^+ ions).

5.1.2 Charge transport calculations

Phase coherent transmission of electrons from one contact to the other through the DNA involves using the Hamiltonian from the DFT calculations discussed in Section 5.1.1 (called the coherent case). In the coherent case, the quantum mechanical phase of the electron evolves as per the single-particle Schrödinger equation, and the electron does not feel the influence of the other

degrees of freedom such as lattice vibrations and solvent environment. We know from prior work that the coherent case yields very low values of the transmission compared to experiments. Decoherence, which represents the interaction of the effective single-particle Hamiltonian with other degrees of freedom, helps us move closer to explaining a set of experiments [132]. Thus, here, we present results for the decoherent case. For the comparison between coherent and decoherent cases, we refer readers to Section 5.2.4.

The charge transport calculations are carried out using the Green's function method by closely following the method used in Refs. [132] and [137]. To model decoherence, we used decoherence probes at each atom in the system [137]. Our primary constraint is setting the net current at each probe equal to zero. The electrical contacts are made at the cytosine bases in the 3′-end and 5′-end to mimic an experimental configuration (see Figure 5-2).

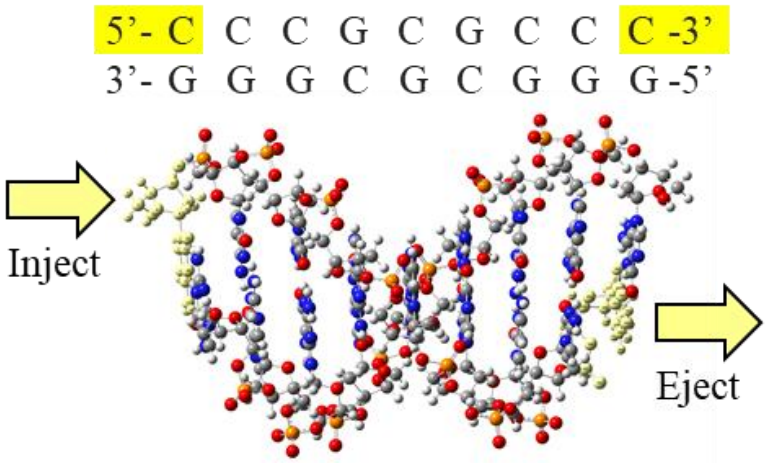


Figure 5-2. The sequence and the atomic structure of the B-DNA strand. The yellow arrows represent the contact and open boundary conditions. The yellow highlight atoms on the two ends are where the contact self-energies are applied.

To obtain the Fock (H_0) and overlap (S_0) matrices from the DFT calculations (Section 5.1.1), we set the dielectric constant to be 78.3553 and 1.0 for wet and dry conditions, respectively via the PCM model. Using Löwdin transformation, the nonorthogonal basis set Fock matrix H_0 is converted to an orthogonal basis set Hamiltonian H:

$$H = S_0^{-\frac{1}{2}} H_0 S_0^{-\frac{1}{2}} \tag{5-1}$$

The diagonal elements of H represent the energy levels at each atomic orbital of the system. The off-diagonal elements of H represent the coupling between the different atomic orbitals. With energy levels and coupling in place, we used the Green's function method; in particular, the retarded Green's function (G^r), is calculated using

$$[E - (H + \Sigma_L + \Sigma_R + \Sigma_D)]G^r = I, \tag{5-2}$$

where E is the energy and H is the Hamiltonian. $\Sigma_{L(R)}$ is the left (right) contact retarded self-energy, which represents the coupling strength of the contacts to the DNA molecule. The self-energy Σ_D is due to the decoherence probes. Using the wide-band limit approximation [138], we assume an energy-independent self-energy, which is defined as

$$\Sigma_{L(R)} = -i\Gamma_{L(R)}/2, \tag{5-3}$$

where i is the imaginary unit, and $\Gamma_{L(R)}$ the coupling strength between the DNA and the left (right) contact.

The self-energy due to the phase-breaking probe (Σ_D) represents the influence of these probes on the DNA. They are defined in a similar manner:

$$\Sigma_D = -\sum_j \frac{i\Gamma_j}{2}. \tag{5-4}$$

The summation over j on the right-hand side is over the probes. Γ_j represents the coupling strength between the probe and the DNA is taken as an energy-independent parameter.

We set the left (right) contact scattering rate $\Gamma_{\mathrm{L}}(\Gamma_{\mathrm{R}})$ to 100 meV and the decoherence scattering rate to 10 meV, which are within the acceptable range [132], [139]. The temperature is assumed to be 298K. The current at the i^{th} probe is defined as

$$I_i = \frac{2q}{h}\sum_{j=1}^{N}\int_{-\infty}^{+\infty} T_{ij}(E)\big[f_i(E) - f_j(E)\big]dE \tag{5-5}$$

$$I_i = \frac{2e}{h}\sum_{j=1}^{N} T_{ij}\big(\mu_i - \mu_j\big),\ i = 1,2,3,\ldots,N, \tag{5-6}$$

where $T_{ij} = \Gamma_i G^r \Gamma_j G^a$ is the transmission probability between the i^{th} and j^{th} probes, $G^a = (G^r)^\dagger$ is the advanced Green's function, and $f_i(E) = \left(1 + \exp\left(\frac{E-E_{fi}}{kT}\right)\right)^{-1}$ is the Fermi distribution. N is the total number of probes in the system (including the contact atoms) and $N_{contacts}$ is the total number of probes on contact atoms.

To ensure current continuity, the current at each probe with respect to energy is set to zero. In our calculations, we applied the decoherence probes at each atom; thus, we have the number of probes $N_b = N - N_{contacts}$. This condition yields N_b independent equations that help derive the following relation [140].

$$\mu_i - \mu_L = \left(\sum_{j=1}^{N_b} W_{ij}^{-1} T_{jR}\right)(\mu_R - \mu_L),\ i = 1,2,3,\ldots,N_b, \tag{5-7}$$

where W_{ij}^{-1} is the inverse of $W_{ij} = (1 - R_{ii})\delta_{ij} - T_{ij}\big(1 - \delta_{ij}\big)$, and R_{ii} is the reflection probability at probe ^{i}, and is given by $R_{ii} = 1 - \sum_{i\neq j}^{N} T_{ij}$. Further, since the current at the left and right contacts is not zero, we can write the equation for current as

$$I_L = -I_R = \frac{2e}{h} T_{eff}(\mu_L - \mu_R). \quad (5\text{-}8)$$

Comparing eq. (5-6) to eq. (5-8), we obtain the effective transmission:

$$T_{eff} = T_{LR} + \sum_{i,j=1}^{N_b} T_{Li} W_{ij}^{-1} T_{jR}, \quad (5\text{-}9)$$

where T_{LR} is the coherent transmission from the left to the right contact, and the second term is the decoherent transmission component.

5.2 Results and Discussion

To study the role of static counterions (Na^+) and solvent in affecting charge transport, we start by investigating the water Na^+ and dry Na^+ cases. We use the PCM with the dielectric constant (ε) to model the solvent. We first observe that the band gap significantly depends on the dielectric constant. For the dry Na^+ case, the band gap is 0.77 eV, much smaller than the value of 4.03 eV for the water Na^+ case (see the first row of Table 5-1).

Table 5-1: HOMO band, LUMO band, and Band gap comparisons between four different cases.

		Water (ε = 78.3553)		Dry (ε = 1)	
		Removed Na^+	Na^+	Removed Na^+	Na^+
Band gap (eV)		4.0795	4.0333	0.4560	0.7714
HOMO Band	HOMO (eV)	-5.0708	-5.0771	-2.1045	-3.4365
	Location	Guanine	Guanine	Guanine	Guanine
	Levels	9	9	10	9
	Bandwidth (eV)	0.3941	0.3927	1.3323	1.0344
	Level density (eV^{-1})	22.8368	22.9183	7.5058	8.7007
LUMO Band	LUMO (eV)	-0.9913	-1.0438	-1.6485	-2.6651
	Location	Cytosine	Cytosine	Na^+	Na^+
	Levels	9	9	15	16
	Bandwidth (eV)	0.2204	0.2468	1.3353	1.2952
	Level density (eV^{-1})	40.8348	36.4668	11.2334	12.3533

The transmission is a measure of electron flow between the left and right contacts through the DNA (Figure 5-2). As discussed in the Methods Section 5.1, the electrons interact with the decoherence probes as they flow through the DNA. It has been previously shown that the conductance of DNA molecules can be altered by conformation and solvent environments [15], [44], [45], [46]. However, their individual effect in determining the conductance is hard to distinguish in prior modeling studies. Our goal is to systematically understand how counterions and solvent dielectric individually influence the transmission. Therefore, we keep the DNA

coordinates or geometry fixed throughout our calculations, and we compare our results for transmission obtained from the same DNA structure with counterions in the dry and water environments.

5.2.1 Role of counterions and solvent dielectric

Figure 5-3 shows that the transmission of the water Na^+ case at the HOMO and LUMO bands is primarily through the atoms of the DNA bases rather than Na^+ ions. This is further enunciated by plotting the wave functions of the highest nine HOMO and the lowest nine LUMO energy levels in Figure 5-3 (b) and (c). The wave functions at these energy levels all lie on the guanines (HOMO band) and cytosines (LUMO band). A previous computational study [141] used structures with Na^+ ions placed further away than in our calculations (2-3 Å from the phosphate group), and they found that the HOMO and LUMO levels are determined by the DNA bases, similar to our results. In Figure 5-3 (a), we note that the band of energies around 0 eV is due to electron transport along the Na^+ ions since the wave functions of these energy levels are primarily localized on Na^+ ions.

For the dry Na^+ case, we also observe that the transmission at the HOMO band is through the DNA molecule (see Figure 5-4). Figure 5-4 (b) shows that the highest nine HOMO band wave functions lie on the guanines. In contrast, the transmission at the LUMO band present around energies of -2 eV is due to Na^+ ions. The wave functions corresponding to the lowest sixteen LUMO energies all lie on the Na^+ ions, as shown in Figure 5-4 (c). The transmission resonances at the unoccupied orbitals around $-0.8\ eV$ are due to transport through both the DNA molecule and the Na^+ ions because the wave functions at these energies are localized on either cytosines or Na^+ ions.

Although the actual HOMO energy level is essential, the energy separation between molecular orbitals and their spatial distribution is also critical for the electrons to hop from one energy level to another, which leads to better transmission. For easier comparison, we define the number of energy levels divided by the Bandwidth as *level density*. We observe that the level density depends on the dielectric constant from the results summarized in Table 5-1. In general, the water cases have a higher level density than the dry cases, and thus larger transmission. Similarly, LUMO bands have higher level density than HOMO bands, and thus larger transmission. Although the transmission is not linearly proportional to level density, their correlation is not neglectable.

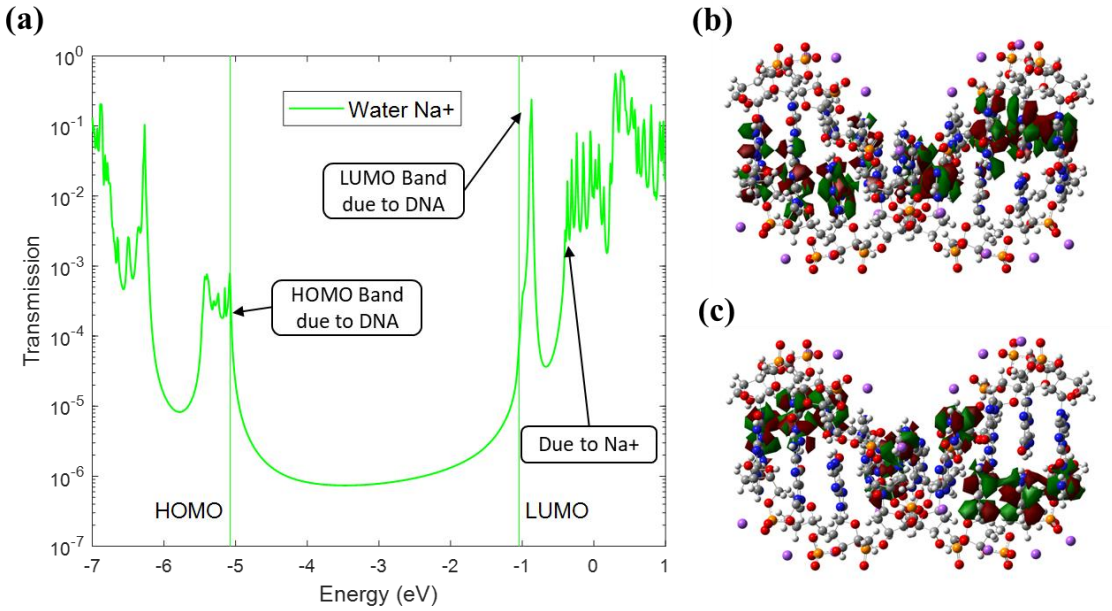


Figure 5-3. (a) Decoherent transmission of DNA with Na^+ ions in water environment (water Na^+). The arrowed-boxes indicate the localization of several energy levels based on wave function plots. (b) The wave functions of the highest nine HOMO energy levels (HOMO band) are localized on guanine bases. (c) The wave functions of the lowest nine LUMO energy levels (LUMO band) are localized on cytosine bases.

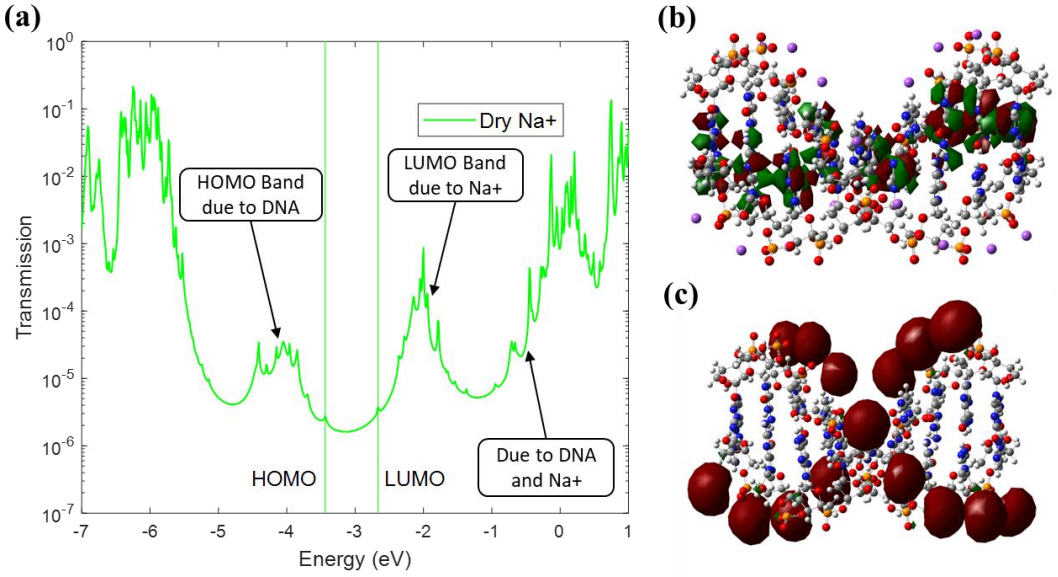


Figure 5-4. (a) Decoherent transmission of DNA with Na^+ ions in a dry environment (dry Na^+). The arrowed- boxes indicate the localization of several energy levels based on wave function plots. (b) The wave functions of the highest nine HOMO energy levels (HOMO band) are localized on guanine bases. (c) The wave functions of the lowest sixteen LUMO energy levels (LUMO band) are localized on Na^+ ions. The energies above the LUMO band (-1 eV onwards) consist of energy levels on the cytosine and Na^+ ions (not shown here). See Section 5.2.4 for coherent results, which lend further support to these observations.

To further investigate the role of counterions, we performed a simulation by randomly removing one of the 16 Na^+ counterions from the DNA model and setting the system's total charge to a negative one. For instance, by removing the seventh Na^+ ion from the water Na^+ case, no significant difference in transmission occurs at the HOMO and LUMO energies, as seen in Figure 5-5 (a). Only the transmission peaks at energies above the LUMO band become smaller. This characteristic, besides the wave function, provides evidence that the band of energy levels around 0 eV is formed by Na^+ ions. The level densities of both HOMO and LUMO bands stay relatively constant with or without removing one of the 16 Na^+ ions.

In comparison, removing the seventh Na^+ ion from the dry Na^+ case lowers the transmission of the LUMO band significantly in Figure 5-5 (b). Missing one Na^+ ion breaks the transport pathway of electrons, which is clearly built on 16 Na^+ ions. While the LUMO Bandwidth stays relatively constant, the level density decreases by about 1.1 eV^{-1} after losing one energy level. Therefore, in Figure 5-5 (b), the LUMO levels of both dry cases are aligned for easy comparison. Meanwhile, we noticed HOMO-1 of the dry removed Na^+ case coincides with the HOMO of the dry Na^+ case. A higher occupied energy level appears in the dry removed Na^+ case beyond the original HOMO band of the dry Na^+ case. The location of this new HOMO energy level further decreases the band gap to 0.46eV. One extra energy level in the HOMO band causes its Bandwidth to extend by about 0.3 eV and its level density to decrease by about 1.2 eV^{-1}. As a result, the average transmission value across the HOMO band is slightly lower than the dry Na^+ case.

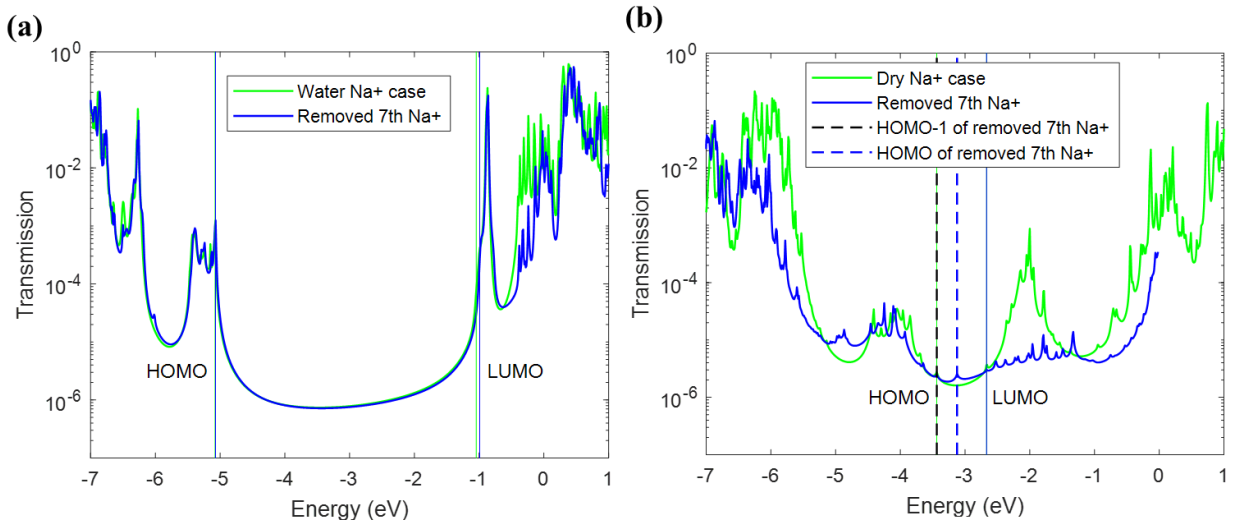


Figure 5-5. Comparison of the transmission plots with (green) and without (blue) the seventh Na^+ ion removed. All other counterions are present. (a) No significant difference is seen with the water solvent. (b) A significant difference is observed for dry cases at both the HOMO and LUMO bands. For subplot (b), we have shifted the energy to align the LUMO level of both molecules for easy comparison.

5.2.2 Hopping parameters between neighboring bases

To understand the underlying reasons for the above observations on the role of the solvent and counterions in transport properties, we analyze the width of the HOMO and LUMO bands, the on-site potential at the bases, and the hopping strength between neighboring bases. For this, we first arrange the Hamiltonian H [from eq. (5-1)] in the order of DNA bases,

$$H = \begin{bmatrix} H_{11} & H_{12} & \cdots & H_{1N} \\ H_{21} & H_{22} & & \vdots \\ \vdots & & \ddots & \\ H_{N1} & \cdots & & H_{NN} \end{bmatrix} \tag{5-10}$$

where N is the number of bases ($N = 18$). The diagonal subblocks H_{ii} correspond to the Hamiltonian of base i, and the off-diagonal subblock H_{ij} represents the coupling between bases i and j. Then we perform the following transform to diagonalize all diagonal subblocks of H,

$$\widehat{H} = U^{\dagger} H U \tag{5-11}$$

where U is a block diagonal matrix. To obtain U, we calculate the eigenvectors of each subblock of H, then construct the entire U matrix,

$$U = \begin{bmatrix} U_{11} & 0 & \cdots & 0 \\ 0 & U_{22} & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & U_{NN} \end{bmatrix} \text{ and } U_{ii} = eigenvectors(H_{ii}) \tag{5-12}$$

The resulting Hamiltonian $\widehat{H}$ is similar in form to eq. (5-10). The diagonal subblocks of $\widehat{H}$ correspond to the energy levels of each DNA base, and the off-diagonal subblocks represent the hopping strength between energy levels on two different bases. Finally, in our discussion below, we restrict ourselves to one energy level (such as HOMO or LUMO).

The HOMO band of the water case lies from -5.47 to -5.08 eV, corresponding to a Bandwidth of 0.39 eV (see Table 5-1). In comparison, the HOMO level density of the dry case is about 2.6 times smaller. The underlying reason for this is the low dielectric constant in the dry case. It results in a larger separation (smaller level density) between the on-site energy levels at the bases corresponding to the HOMO band. On the other hand, the high dielectric constant of water makes the on-site energies corresponding to the HOMO band energetically closer (larger level density).

The on-site potential of the bases corresponding to the HOMO band in the water and dry cases are shown in Figure 5-6. Although the hopping terms are similar in the two cases (especially coupling between G-G neighbor bases), the on-site potentials of the water case are more uniform and closer together. This behavior can also be seen by comparing their coefficients of variation which are -0.0581 (water case) vs -0.1811 (dry case). The HOMO level density of the water case is 22.92 eV^{-1}, which is about 2.6 times larger than the dry case (8.70 eV^{-1}). As a result, the average transmission value across the HOMO band is approximately one to two orders

of magnitude larger in the water case as opposed to the dry case. Therefore, the electrons travel through the DNA more efficiently in the water case.

The on-site potentials of the bases corresponding to the LUMO band in the water and dry cases are shown in Figure 5-7. Unlike all other bands, the LUMO band of the dry Na^+ case comprises 16 energy levels localized on the 16 Na^+ ions instead of the DNA molecule. The on-site potentials of the Na^+ ions in the LUMO band range in energies from -2.43 to -1.45 eV. The hopping terms between Na^+ ions in the LUMO band are primarily in the range of 50-70 meV.

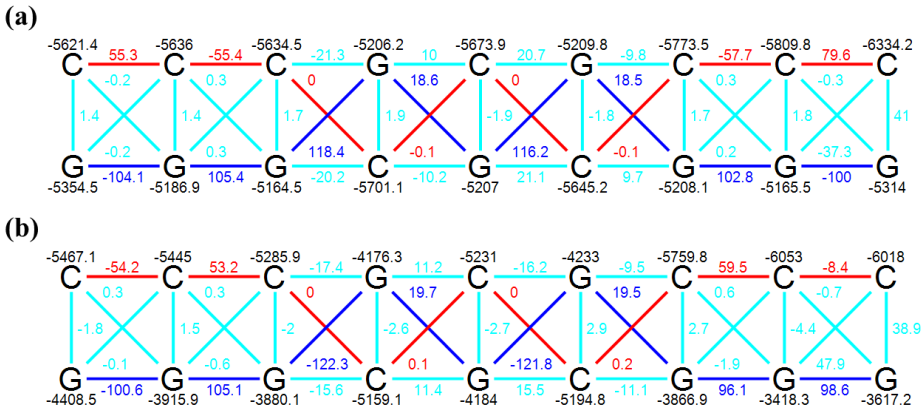


Figure 5-6. The hopping parameters between neighboring bases at (a) the highest 18 HOMO levels of the water Na^+ case and (b) the highest 18 HOMO levels of the dry Na^+ case. Units are in meV. For (a) and (b), the coefficient of variation of the on-site potentials is -0.0581 and -0.1811, respectively.

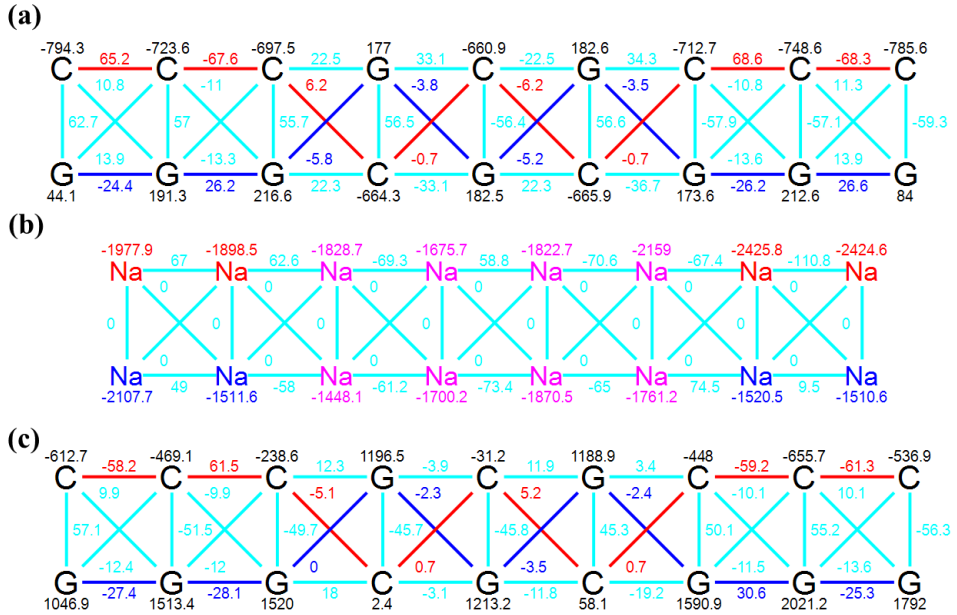


Figure 5-7. The hopping parameters between neighboring bases at (a) the lowest 18 LUMO levels of the water Na^+ case, (b) the lowest 16 LUMO levels of the dry Na^+ case, and (c) the even higher LUMO levels (LUMO + 16 ~ LUMO + 33) of the dry Na^+ case. Units are in meV.

5.2.3 Larger basis set

In the context of the level of theory for the density functional theory (DFT) calculations, the B3LYP/6-31G(d,p) has been used to calculate the ionization potential of nucleobases extensively. These calculations find the B3LYP/6-31G(d,p) to yield the correct trend in ionization potential as experiments but with an offset in values of about 300 meV. Other more expensive methods, such as CCSD, MP2, and cc-pVTZ, may give a higher accuracy for the ionization potential, but what makes B3LYP/6-31G(d,p) a method of choice is a balance between calculation accuracy and reasonable computational cost. Since our system under study comprises a relatively large molecular structure (~700 atoms), we need to use the entire Hamiltonian and Overlap matrices to calculate the transport properties.

Table 5-2. Number of functions with various basis sets. Using the same 9 base pair DNA.

		H - He	Li - Ne	Na - Ar	# of functions
Pople basis sets (s=1, p=3, d=5, d*=6)	6-31G	[2*s*] → 2	[3*s*2*p*] → 9	[4*s*3*p*] → 13	3943
	6-31G(d,p)	[2*s*1*p*] → 5	[3*s*2*p*1*d**] → 15	[4*s*3*p*1*d**] → 19	6823
	6-311G(d,p)	[3*s*1*p*] → 6	[4*s*3*p*1*d*] → 18	[6*s*5*p*1*d*] → 26	8290
Correlation-consistent basis sets (s=1, p=3, d=5, f=7)	cc-pVDZ	[2*s*1*p*] → 5	[3*s*2*p*1*d*] → 14	[4*s*3*p*1*d*] → 18	6444
	cc-pVTZ	[3*s*2*p*1*d*] → 14	[4*s*3*p*2*d*1*f*] → 30	[5*s*4*p*2*d*1*f*] → 34	14326

* The 6 d-type polarization function is added to 6-31G set, while only 5 to 6-311G. For both 6-31G and 6-311G sets, f-type functions are added in groups of 7.

To demonstrate that our results are more or less basis-set-independent, we performed the same simulation procedure as the one discussed in Section 5.1 on the sequence of 5'-CCCGCGCCC-3' with three different basis sets for DFT calculation: (a) B3LYP/6-311G(d,p), (b) B3LYP/cc-pVDZ, and (c) B3LYP/cc-pVTZ (see Table 5-2). Please note that B3LYP/6-311G(d,p) and B3LYP/cc-pVTZ are larger basis sets than B3LYP/6-31G(d,p). To evaluate the robustness, we focused on B3LYP/cc-pVTZ as a benchmark since it has been shown to yield more accurate results but with an increase in computational time. The transmission and wave function plots calculated using B3LYP/cc-pVTZ are shown in the following figures. A comparison between calculations (Figure 5-3 vs. Figure 5-8 and Figure 5-4 vs. Figure 5-9) indicates that using the B3LYP/cc-pVTZ basis set produces results consistent with the B3LYP/6-31G(d,p) basis set. Analysis of the LUMO orbitals under dehydrated conditions (see Figure 5-9)

shows that orbitals are still localized on Na^+ ions when the cc-pVTZ basis set is used. Similarly, in the water solvent (as shown in Figure 5-8), HOMO orbitals continue to be localized on Guanine bases. In addition, the transmission and wave function plots, calculated using B3LYP/6-311G(d,p) and B3LYP/cc-pVDZ, show the same behavior (see Figure 5-10, Figure 5-11, Figure 5-12, Figure 5-13). Therefore, we anticipate the B3LYP/6-31G(d,p) basis set provides a balance between accuracy and computational cost and the conclusions drawn in the paper are accurate enough to be considered as robust and independent of the choice of basis sets.

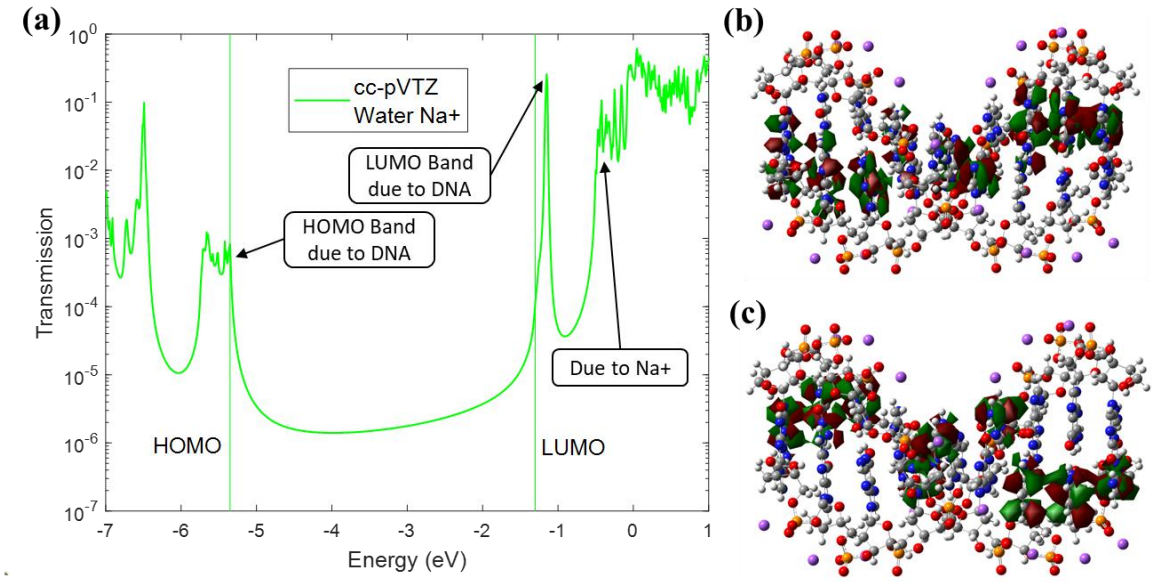


Figure 5-8. DFT calculations with B3LYP/cc-pVTZ basis set for (a) Decoherent transmission of DNA with Na^+ ions in water environment (*Water* Na^+). The arrowed-boxes indicate the localization of several energy levels based on wavefunction plots. (b) The wavefunctions of the highest nine HOMO energy levels (HOMO band) are localized on Guanine bases. (c) The wavefunctions of the lowest nine LUMO energy levels (LUMO band) are localized on Cytosine bases.

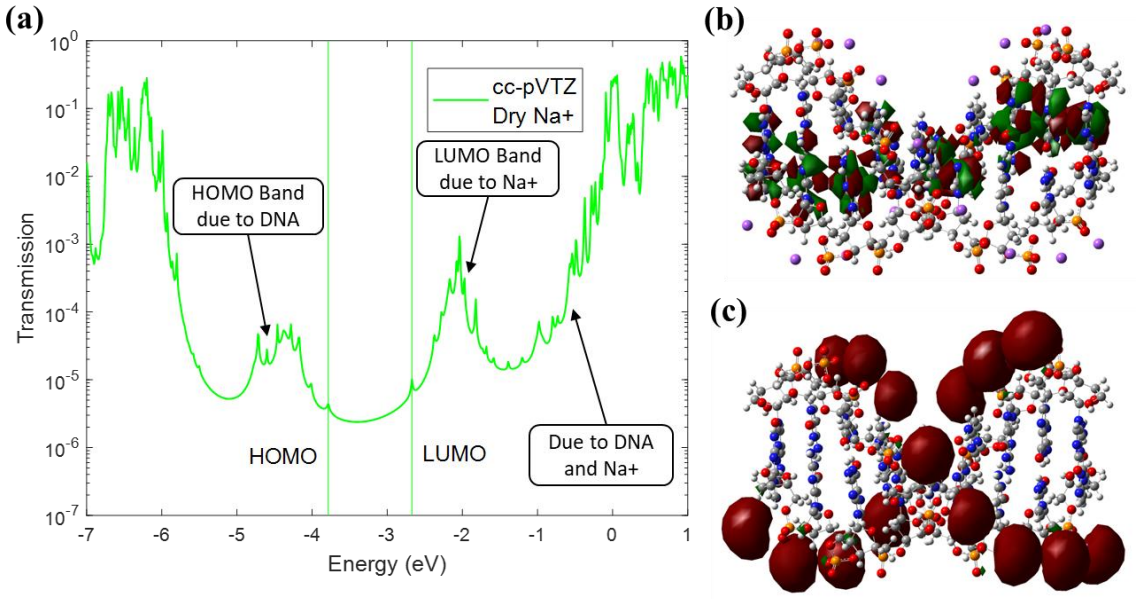


Figure 5-9. DFT calculations with B3LYP/cc-pVTZ basis set for (a) Decoherent transmission of DNA with Na^+ ions in a dry environment (*Dry* Na^+). (b) The wavefunctions of the highest nine HOMO energy levels (HOMO band) are localized on Guanine bases. (c) The wavefunctions of the lowest sixteen LUMO energy levels (LUMO band) are localized on Na^+ ions.

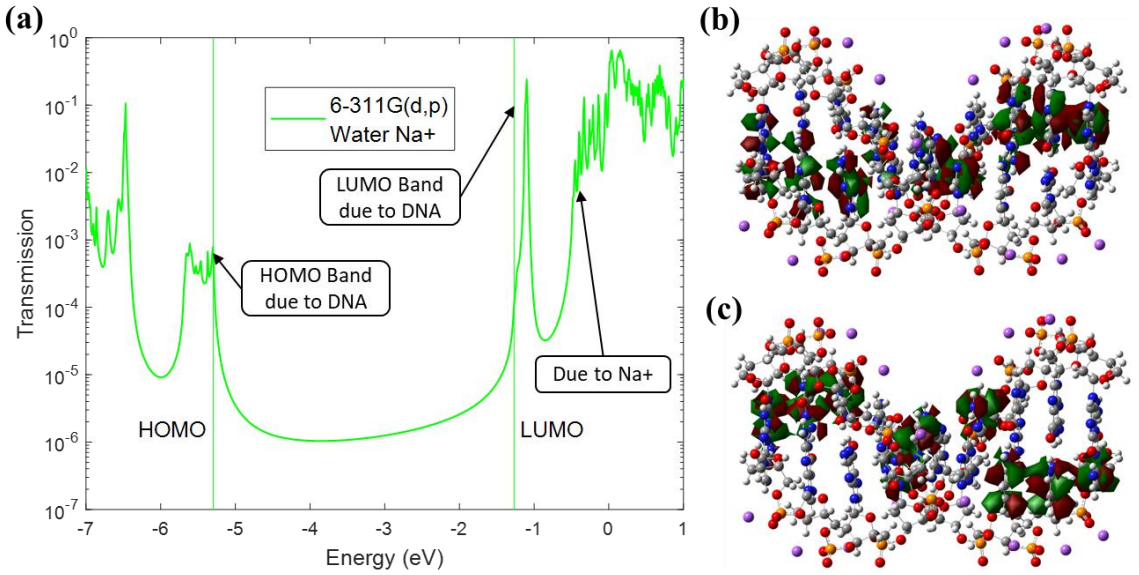


Figure 5-10. DFT calculations with B3LYP/6-311G(d,p) basis set for (a) Decoherent transmission of DNA with Na^+ ions in a water environment (*Water* Na^+). (b) The wavefunctions of the highest nine HOMO energy levels (HOMO band) are localized on Guanine bases. (c) The wavefunctions of the lowest nine LUMO energy levels (LUMO band) are localized on Cytosine bases.

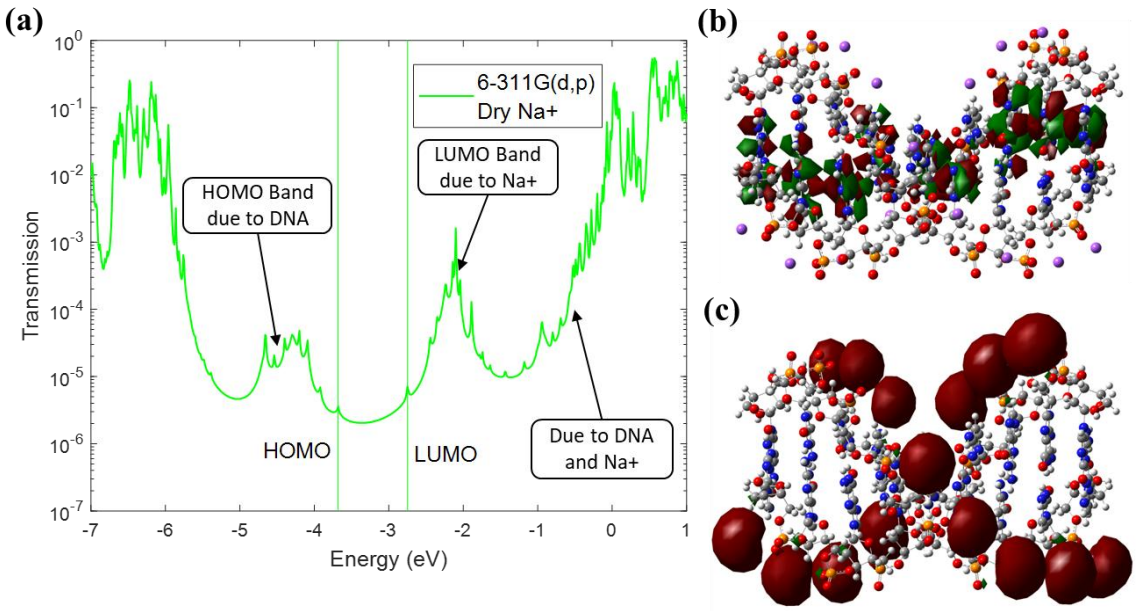


Figure 5-11. DFT calculations with B3LYP/6-311G(d,p) basis set for (a) Decoherent transmission of DNA with Na^+ ions in a dry environment (*Dry* Na^+). (b) The wavefunctions of the highest nine HOMO energy levels (HOMO band) are localized on Guanine bases. (c) The wavefunctions of the lowest sixteen LUMO energy levels (LUMO band) are localized on Na^+ ions.

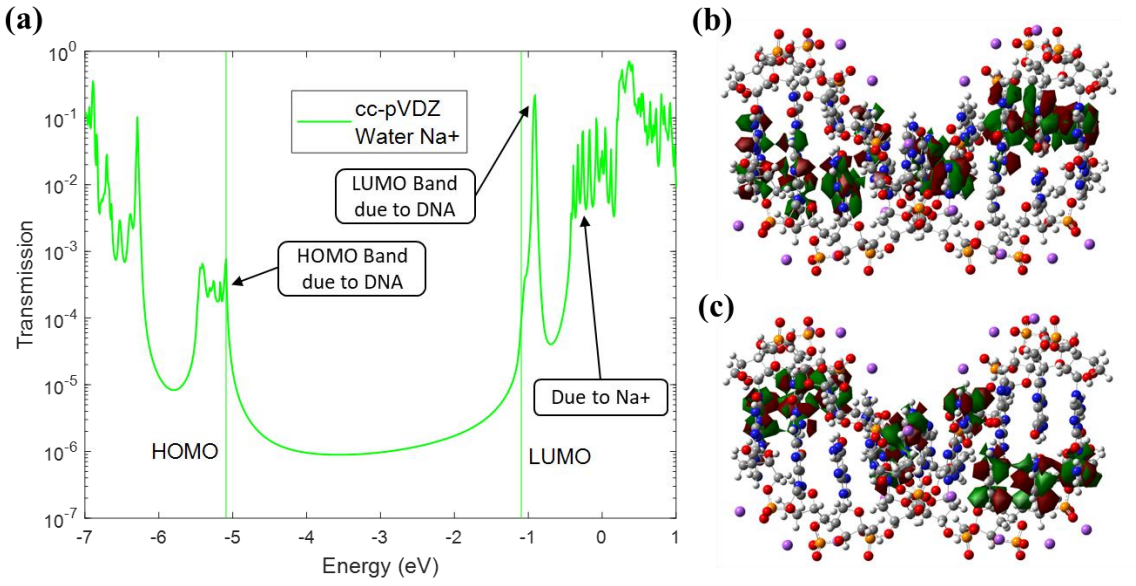


Figure 5-12. DFT calculations with B3LYP/cc-pVDZ basis set for (a) Decoherent transmission of DNA with Na^+ ions in a water environment (*Water* Na^+). (b) The wavefunctions of the highest nine HOMO energy levels (HOMO band) are localized on Guanine bases. (c) The wavefunctions of the lowest nine LUMO energy levels (LUMO band) are localized on Cytosine bases.

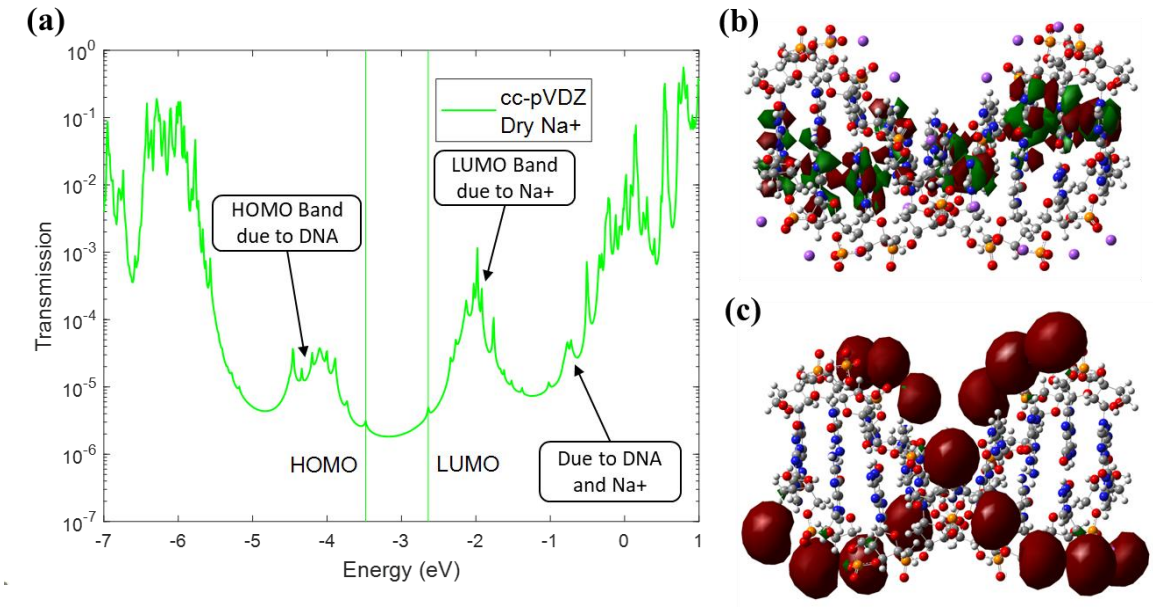


Figure 5-13. DFT calculations with B3LYP/cc-pVDZ basis set for (a) Decoherent transmission of DNA with Na^+ ions in a dry environment (*Dry* Na^+). (b) The wavefunctions of the highest nine HOMO energy levels (HOMO band) are localized on Guanine bases. (c) The wavefunctions of the lowest sixteen LUMO energy levels (LUMO band) are localized on Na^+ ions.

5.2.4 Coherent transmission with no *Na*$^+$ ions

We performed coherent transmission calculations along with decoherent ones which are extensively discussed. As discussed in Section 5.1.2, eq. (5-9) shows that effective (decoherent) transmission is equal to coherent transmission plus the component of decoherence probes. Figure 5-14 shows the *Water* Na^+ case, and Figure 5-15 shows the *Dry* Na^+ case. Both plots show no significant differences between the trend of decoherent and coherent transmissions. Therefore, we can conclude the same discussions and results as the ones in Section 5.2.1. In addition, we add another simulation where all Na^+ ions are removed from the Hamiltonian, called *Water No* Na^+ case and *Dry No* Na^+ case (see Table 5-3), which further supports our observations. Figure 5-14 shows the comparison between *Water* Na^+ case and *Water No* Na^+ case. Figure 5-15 shows the comparison between *Dry* Na^+ case and *Dry No* Na^+ case.

Table 5-3: HOMO band, LUMO band, and Band gap comparisons between four different cases.

		Water (ε = 78.3553)		Dry (ε = 1)	
		No Na^+	Na^+	No Na^+	Na^+
Band gap (eV)		4.0527	4.0333	2.5689	0.7714
HOMO Band	HOMO (eV)	-5.0703	-5.0771	-3.4358	-3.4365
	Location	Guanine	Guanine	Guanine	Guanine
	Levels	9	9	9	9

	Bandwidth (eV)	0.3962	0.3927	1.0307	1.0344
	Level density (eV^{-1})	22.7158	22.9183	8.7319	8.7007
LUMO Band	LUMO (eV)	-1.0176	-1.0438	-0.8669	-2.6651
	Location	Cytosine	Cytosine	Cytosine	Na^+
	Levels	9	9	9	16
	Bandwidth (eV)	0.2309	0.2468	0.7401	1.2952
	Level density (eV^{-1})	38.9779	36.4668	12.1605	12.3533

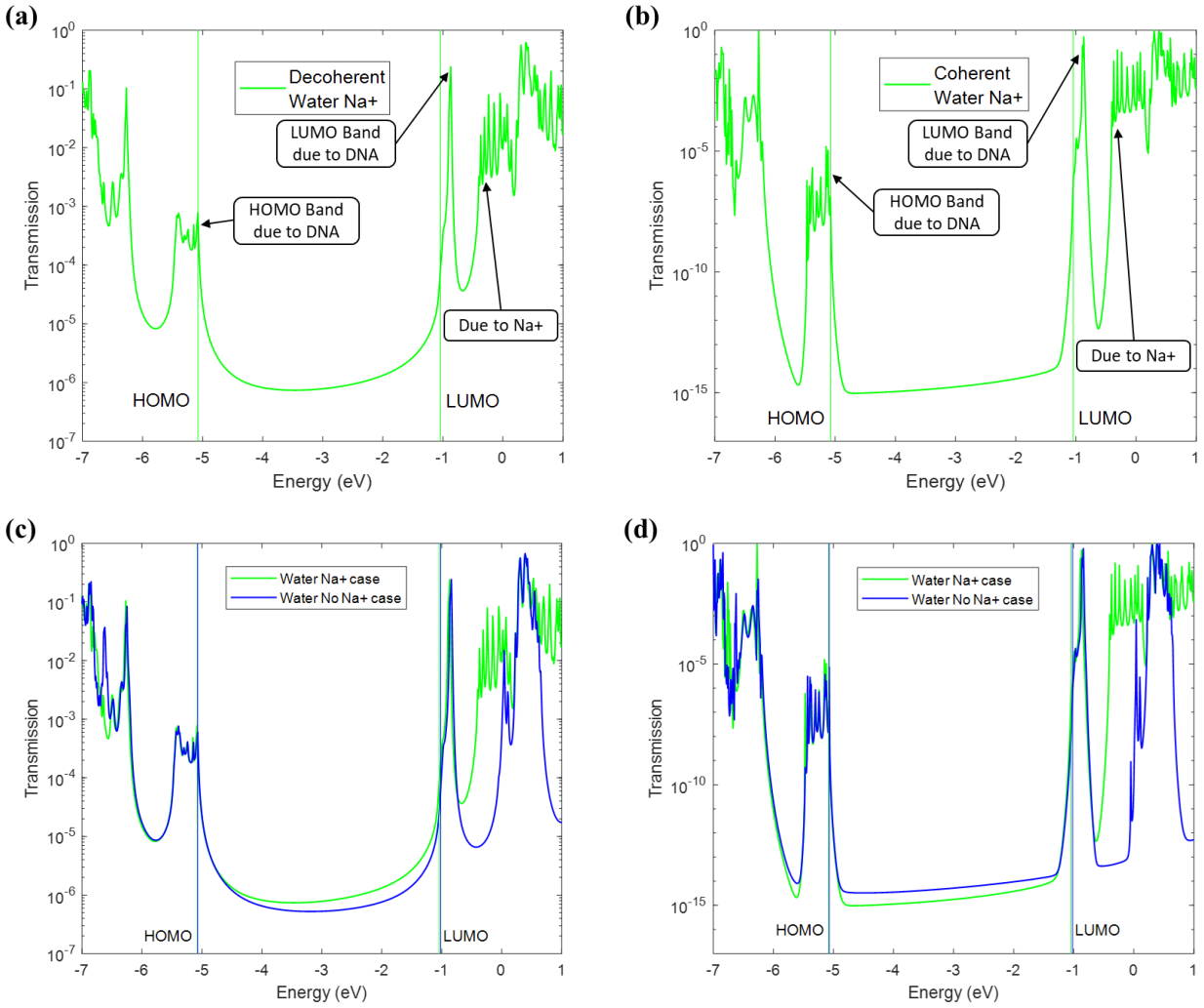


Figure 5-14. Transmission of DNA with Na^+ ions in an implicit water environment. (a) Decoherent transmission, which is the same plot as Figure 5-3 (a). (b) Coherent transmission shows the same trend. (c, d) Comparison between *Water* Na^+ case and *Water No* Na^+ case.

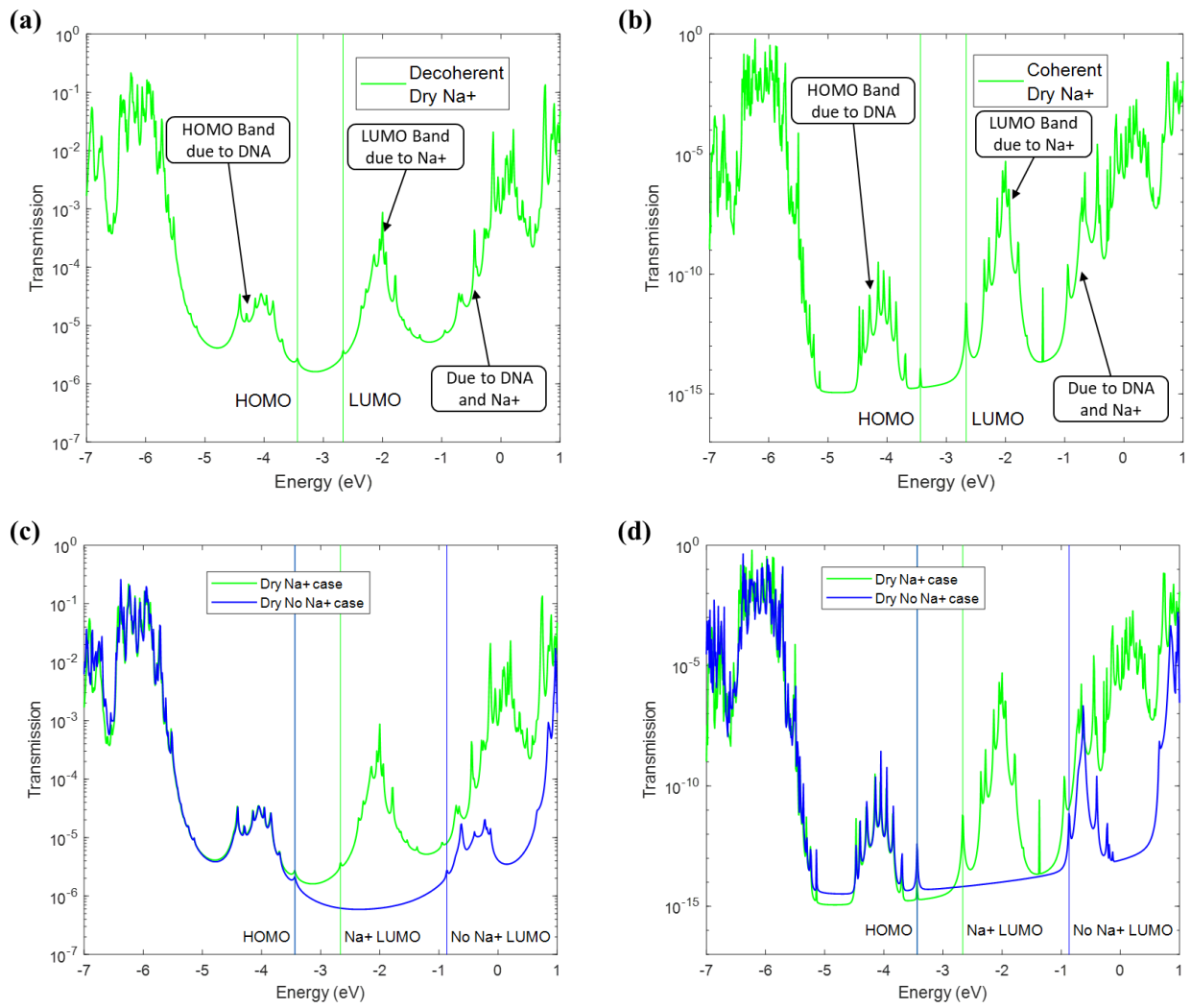


Figure 5-15. Transmission of DNA with Na^+ ions in a dry environment. (a) Decoherent transmission, which is the same plot as Figure 5-4 (a). (b) Coherent transmission shows the same trend. (c, d) Comparison between *Water* Na^+ case and *Water No* Na^+ case.

5.2.5 Results for other DNA sequences

In Section 5.2.1, our discussion focused on one specific nine-base-pair DNA sequence (5'-CCCGCGCCC-3'). However, our conclusions are not limited to this specific sequence. In order to show that our conclusions can apply to other short-strand DNA, we applied the same simulation procedure as the one discussed in Section 5.1 with B3LYP/6-31G(d,p) basis sets for DFT calculation to various DNA strands with different sequences and lengths as shown below:

1. 5'-CCCCCC-3'
   3'-GGGGGG-5' (Figure 5-16 and Figure 5-17)
2. 5'-TTTTTT-3'
   3'-AAAAAA-5' (Figure 5-18 and Figure 5-19)
3. 5'-CCCTTTCCC-3'

3’-GGGAAAGGG-5’ (Figure 5-20 and Figure 5-21)

4. 5’-GGCGCGCGGGCGGGC-3’
   3’-CCGCGCGCCCGCCCG-5’ (Figure 5-22 and Figure 5-23)
5. 5’-GGCGCAAAAACGGGC-3’
   3’-CCGCGTTTTTGCCCG-5’ (Figure 5-24 and Figure 5-25)

We chose these sequences based on previously published work. DNA strands 1~3 come from [141]. DNA strands 4~5 come from [132].

The transmission and wavefunction plots are shown in the figure below. For all investigated sequences, the wavefunctions of the HOMO band are localized on Guanine or Adenine bases. These wavefunctions form a transmission channel from left to right using Guanine or Adenine bases. The channel could utilize both strands of a DNA molecule based on the sequence. The HOMO band’s energy levels are consistent with the number of base pairs. With a water solvent, the wavefunctions of the LUMO band are localized on Cytosine or Thymine bases. Under dehydrated conditions, the wavefunctions of the LUMO band are localized on Na^+ ions. These wavefunctions form two separate channels from left to right (along both backbones of a DNA molecule) using Na^+ ions. The LUMO band’s energy levels are consistent with the number of base pairs or Na^+ ions, respectively. These results show that regardless of the sequence or length, depending on the dielectric constant of the environment, Na^+ ions can significantly impact the charge transport properties of the DNA molecules. Based on the additional calculations, we believe our conclusions can be generalized to other short-strand DNA sequences.

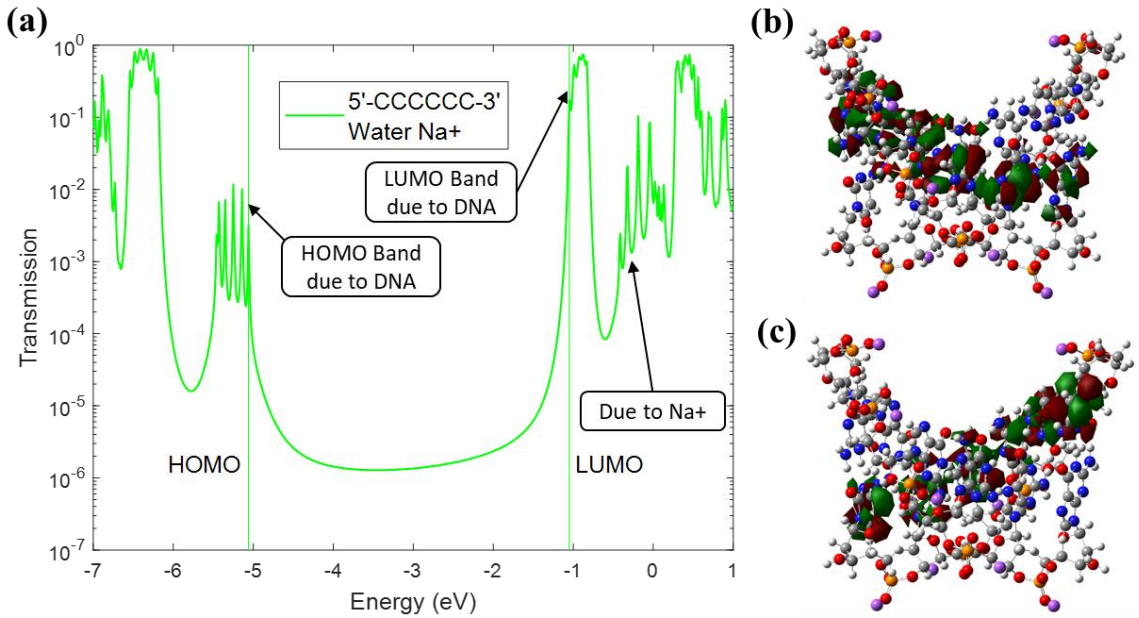


Figure 5-16. DFT calculations with 5’-CCCCCC-3’ sequence for (a) Decoherent transmission of DNA with Na^+ ions in a water environment (*Water* Na^+). (b) The wavefunctions of the highest six HOMO energy levels (HOMO band) are localized on Guanine bases. (c) The wavefunctions of the lowest six LUMO energy levels (LUMO band) are localized on Cytosine bases.

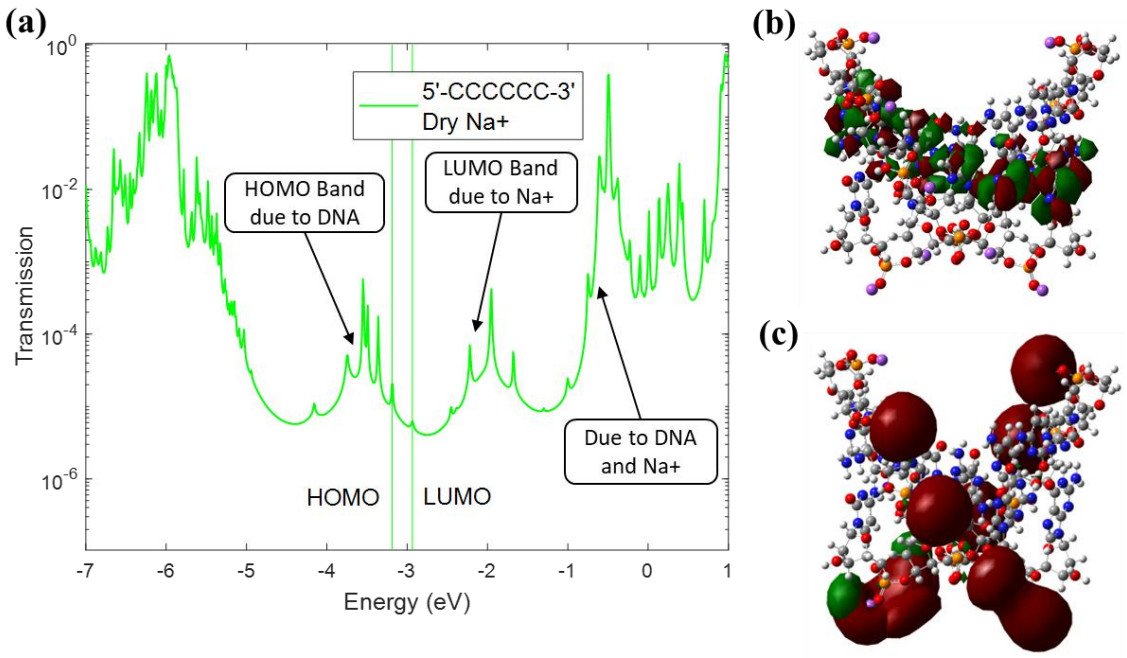


Figure 5-17. DFT calculations with 5'-CCCCCC-3' sequence for (a) Decoherent transmission of DNA with Na^+ ions in a dry environment (*Dry* Na^+). (b) The wavefunctions of the highest six HOMO energy levels (HOMO band) are localized on Guanine bases. (c) The wavefunctions of the lowest ten LUMO energy levels (LUMO band) are localized on Na^+ ions.

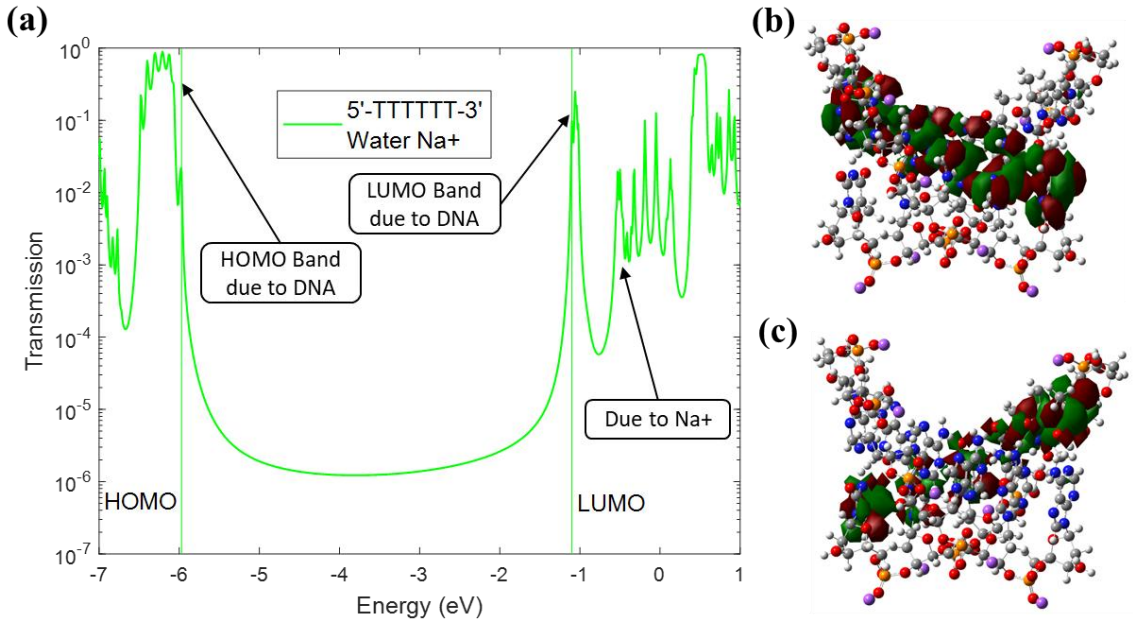


Figure 5-18. DFT calculations with 5'-TTTTTT-3' sequence for (a) Decoherent transmission of DNA with Na^+ ions in a water environment (*Water* Na^+). (b) The wavefunctions of the highest six HOMO energy levels (HOMO band) are localized on Adenine bases. (c) The wavefunctions of the lowest six LUMO energy levels (LUMO band) are localized on Thymine bases.

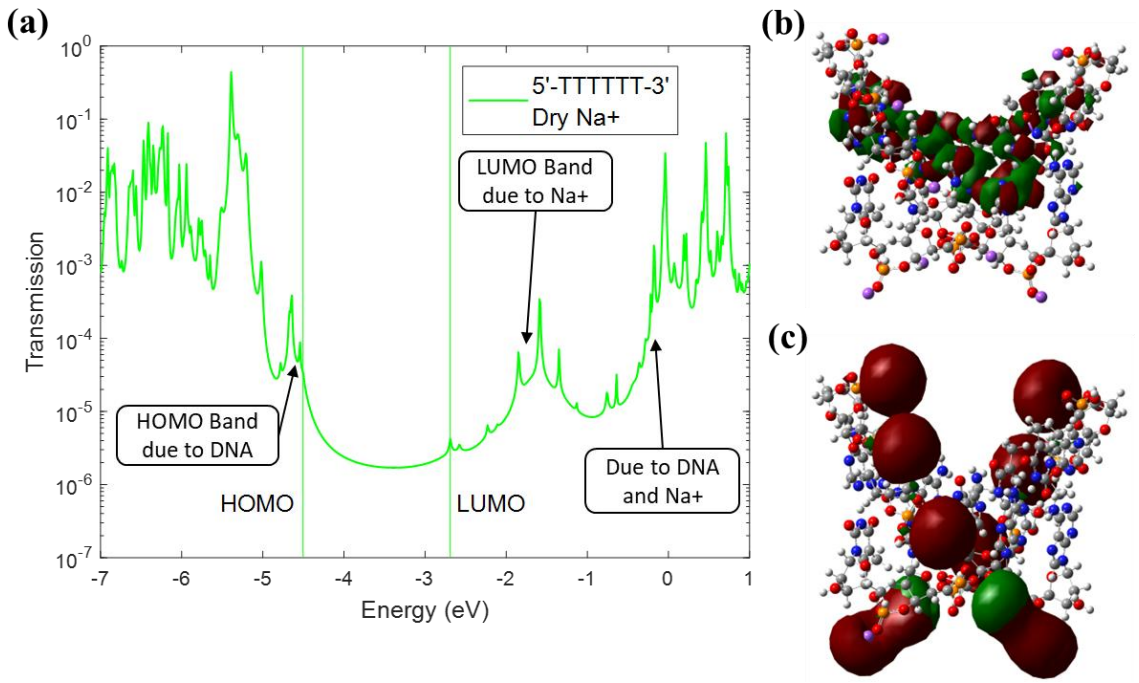


Figure 5-19. DFT calculations with 5'-TTTTTT-3' sequence for (a) Decoherent transmission of DNA with Na^+ ions in a dry environment (*Dry* Na^+). (b) The wavefunctions of the highest six HOMO energy levels (HOMO band) are localized on Adenine bases. (c) The wavefunctions of the lowest ten LUMO energy levels (LUMO band) are localized on Na^+ ions.

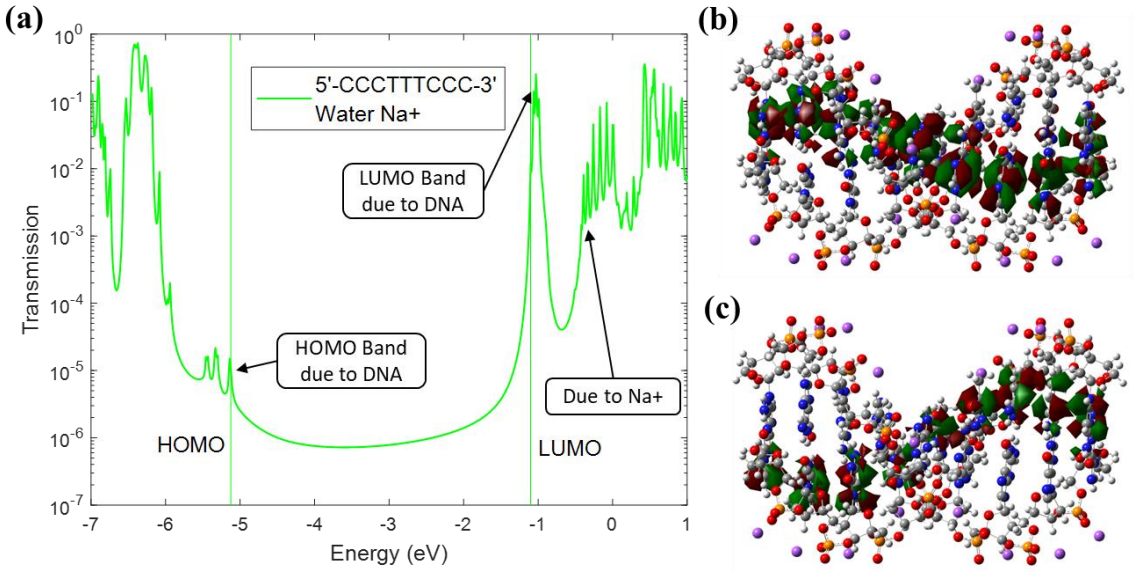


Figure 5-20. DFT calculations with 5'-CCCTTTCCC-3' sequence for (a) Decoherent transmission of DNA with Na^+ ions in a water environment (*Water* Na^+). (b) The wavefunctions of the highest nine HOMO energy levels (HOMO band) are localized on Guanine or Adenine bases. (c) The wavefunctions of the lowest nine LUMO energy levels (LUMO band) are localized on Cytosine or Thymine bases.

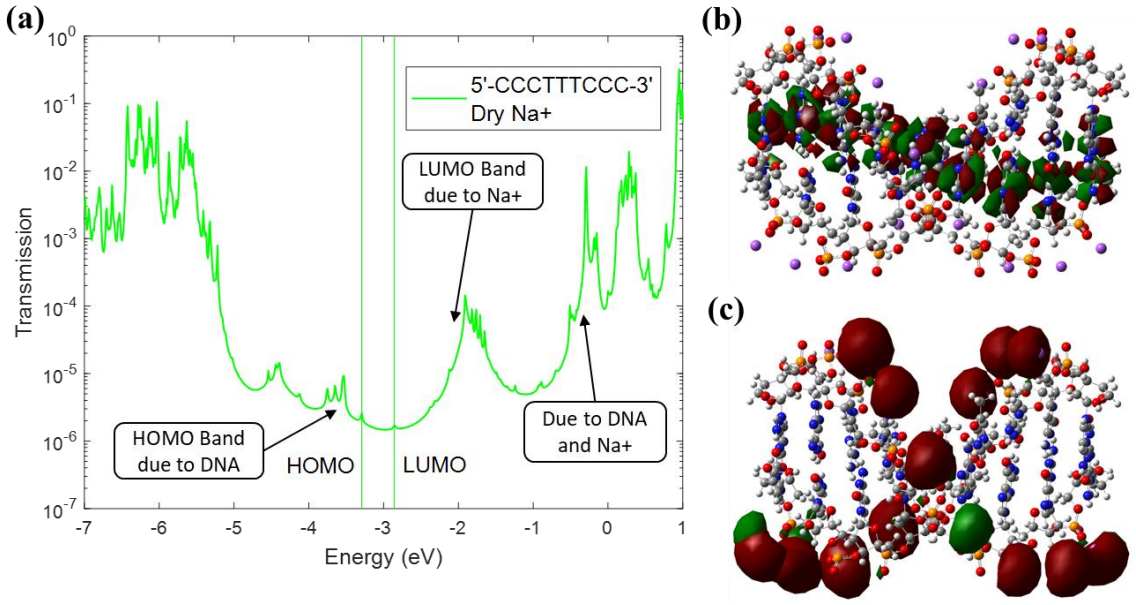


Figure 5-21. DFT calculations with 5'-CCCTTTCCC-3' sequence for (a) Decoherent transmission of DNA with Na^+ ions in a dry environment (*Dry* Na^+). (b) The wavefunctions of the highest nine HOMO energy levels (HOMO band) are localized on Guanine or Adenine bases. (c) The wavefunctions of the lowest sixteen LUMO energy levels (LUMO band) are localized on Na^+ ions.

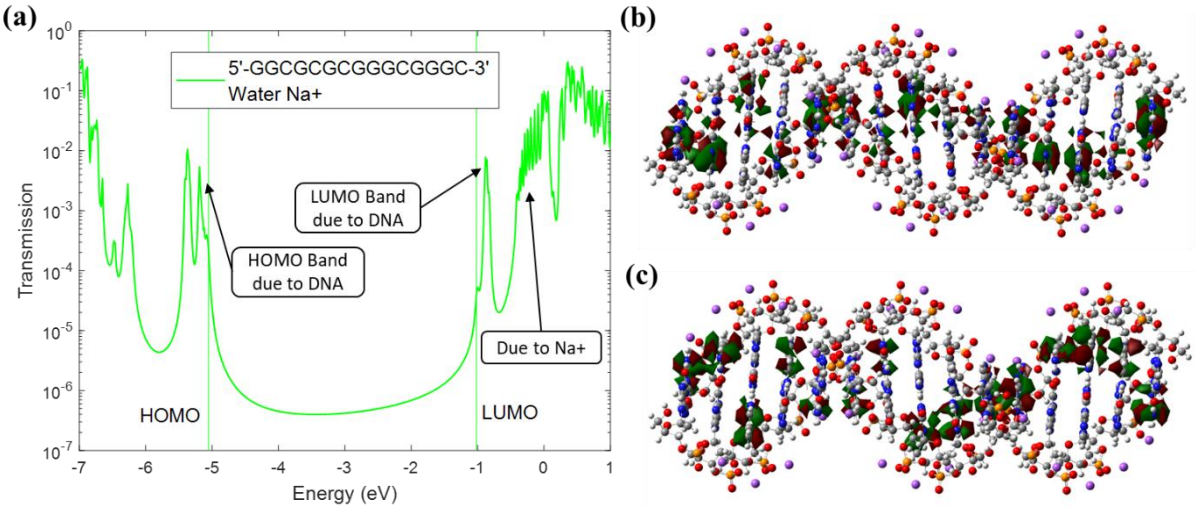


Figure 5-22. DFT calculations with 5'-GGCGCGCGGGCGGGC-3' sequence for (a) Decoherent transmission of DNA with Na^+ ions in a water environment (*Water* Na^+). (b) The wavefunctions of the highest fifteen HOMO energy levels (HOMO band) are localized on Guanine or Adenine bases. (c) The wavefunctions of the lowest fifteen LUMO energy levels (LUMO band) are localized on Cytosine or Thymine bases.

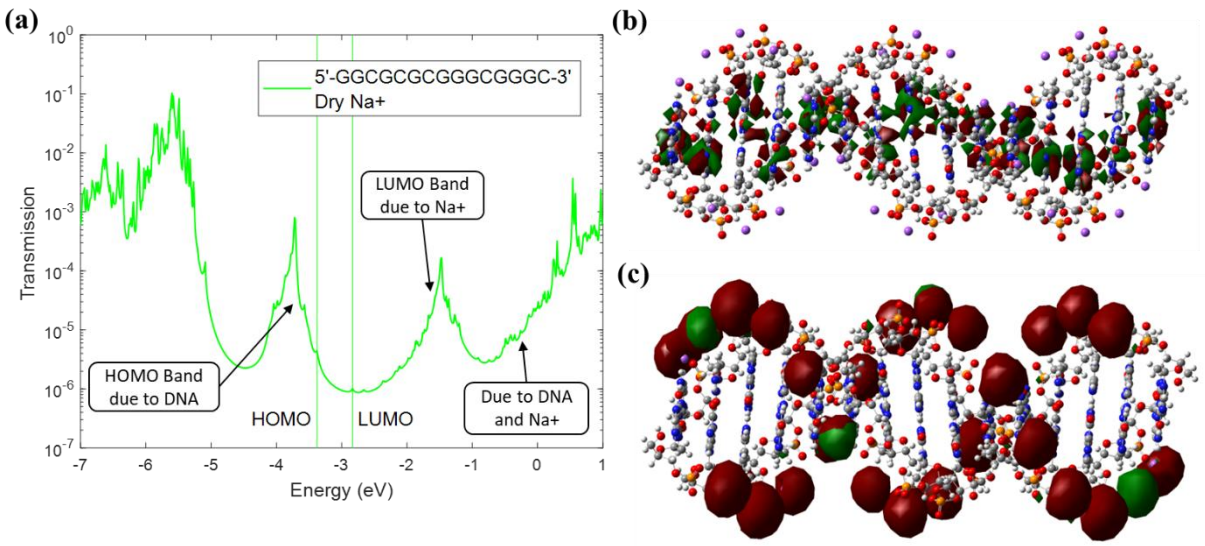


Figure 5-23. DFT calculations with 5'-GGCGCGCGGGCGGGC-3' sequence for (a) Decoherent transmission of DNA with Na^+ ions in a dry environment (*Dry* Na^+). (b) The wavefunctions of the highest fifteen HOMO energy levels (HOMO band) are localized on Guanine or Adenine bases. (c) The wavefunctions of the lowest twenty-eight LUMO energy levels (LUMO band) are localized on Na^+ ions.

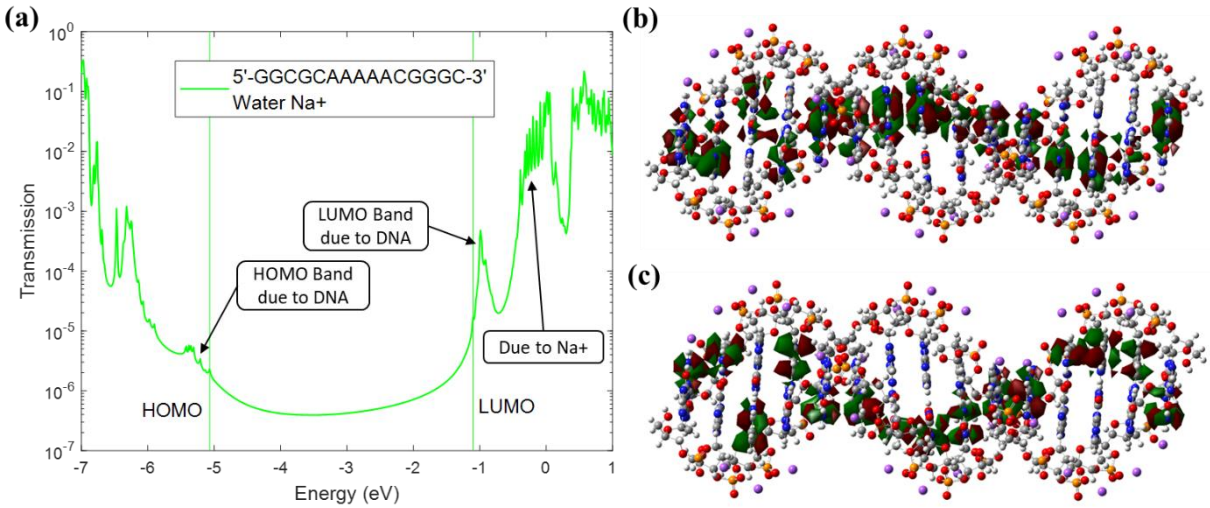


Figure 5-24. DFT calculations with 5'-GGCGCAAAAACGGGC-3' sequence for (a) Decoherent transmission of DNA with Na^+ ions in a water environment (*Water* Na^+). (b) The wavefunctions of the highest fifteen HOMO energy levels (HOMO band) are localized on Guanine or Adenine bases. (c) The wavefunctions of the lowest fifteen LUMO energy levels (LUMO band) are localized on Cytosine or Thymine bases.

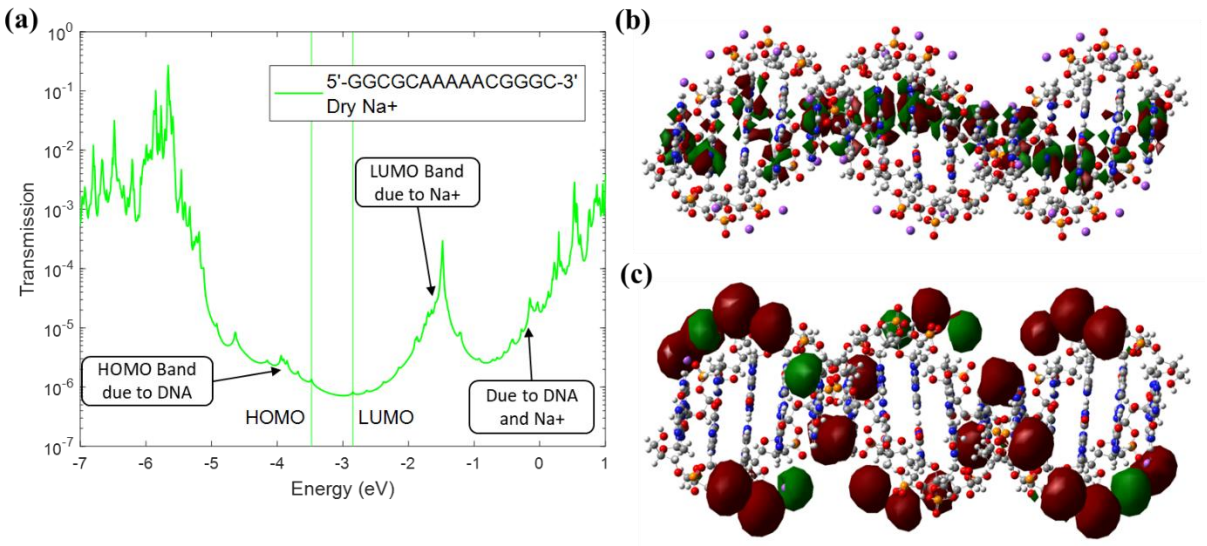


Figure 5-25. DFT calculations with 5'-GGCGCAAAAACGGGC-3' sequence for (a) Decoherent transmission of DNA with Na^+ ions in a dry environment (*Dry* Na^+). (b) The wavefunctions of the highest fifteen HOMO energy levels (HOMO band) are localized on Guanine or Adenine bases. (c) The wavefunctions of the lowest twenty-eight LUMO energy levels (LUMO band) are localized on Na^+ ions.

5.3 Summary

It has been challenging to obtain clear trends in the underlying physics of DNA conduction due to the complexity of environmental conditions. This situation has motivated us to computationally study an important aspect encountered--the effect of the solvent and counterions for a static configuration of atoms. We consider a nine-base-pair double-stranded B-DNA and study the role of counterion arrangement and solvent dielectric constant to determine if there are clear trends in the underlying physics. We use the PCM model for the solvent and consider the dry and fully hydrated environments. By performing calculations on six different DNA sequences, we emphasize the generalizability of the results (additional results are presented in Section 5.2.5).

Depending on the dielectric constant of the surrounding medium, the Na^+ ion is found to significantly impact the charge transport properties of the DNA. From the molecular energy level perspective, Na^+ ions add unoccupied energy levels in the band gap of the DNA in the dry case. On the other hand, the water case adds unoccupied energy levels that have higher energy than the LUMO, which is primarily located on the cytosine bases. Because of the high dielectric constant of water, the interaction between DNA and Na^+ ions is effectively screened. In addition, from the charge transport perspective, the transmission is at least two orders of magnitude larger at the HOMO and LUMO regions of the DNA in the water case than in the dry case. The observed narrower spread of on-site potentials (at the HOMO and LUMO bands) with a water environment supports higher transmission.

In summary, our simulation results demonstrate that it is essential to consider counterions as an individual factor when analyzing the DNA conductance experiments done in the dry case but not necessarily in the water solvent. The higher the dielectric constant, the higher the charge screening effect, thus lowering the coupling between Na^+ ions and DNA molecules. As the presence of Na^+ ions added energy levels within the band gap of the DNA in the dehydrated condition (the dry case), this can further be relevant to utilizing DNA in nanoelectronics applications.

6 Influence of Explicit Water Molecules on the Electronic Properties of DNA

The interaction between DNA molecules and their surrounding environment is crucial in understanding the electronic properties of DNA. In the previous chapters, we performed DFT calculations on a DNA system with the polarizable continuum model (PCM) which tries to capture the screening effect of the solvent without the need to include explicit water molecules [135]. However, during the analysis of DNA conductance traces from SMBJ experiment, we noticed that there is an abnormal conductance peak constantly appearing around $10^{-2}\ G_0$. As suggested by Xiang et al., this conductance peak is likely to be contributed by water molecules [62]. Although we developed statistical methods to remove this peak for better Machine Learning classification results, we believe that it is critical to understand the effect of explicit water molecules from a theoretical perspective.

Since the complexity and computational cost of DFT calculations grow significantly as the system size increases, the majority of DNA electronic properties studies use PCM or other models to mimic the effect of water molecules. Only a few groups have previously tried to capture the effect of water solvent in charge transfer of biomolecules using explicit water molecules in DFT calculations [34], [62], [142], [143], [144], [145], [146], [147], [148]. With less than five explicit water molecules per DNA base pair, several studies have shown that the effect of water is not significant [62], [142], [143], [146]. Wolter et al. showed that transport of electric charge in a micro solvated molecular system will exhibit similar characteristics as in full water solvent, where at least 12 explicit water molecules are needed per DNA base pair [34]. Fukuzawa et al. demonstrated that a solvent shell with explicit water molecules of thickness 8Å is necessary to incorporate the water solvent effect [145]. Westerhoff et al. performed comprehensive calculations by including explicit water molecules within a certain distance away from DNA. They found that 4.5 A and 5.5 A water systems were able to mimic the “infinite” solvation used in the experimental studies [144]. They also compared the difference between explicit water molecules and the PCM model using SemiEmpirical Born-Oppenheimer Molecular Dynamics. They concluded that while the PCM model may not perfectly describe the DNA system in all cases, in most situations, the method does reasonably well [144].

Previous work has shown that including a “sufficient amount” of explicit water molecules could affect the conductivity of DNA molecules. However, there is no uniform conclusion to determine the quantitative amount of water molecules. In this chapter, we perform DFT calculations including explicit water molecules and identify the connection between simulation results and experimental observations.

6.1 Methodology

In a typical MD (Molecular Dynamics) simulation, a DNA strand is solvated with thousands of water molecules and is accompanied by counterions to neutralize its negative charge. Using the TIP3P (transferable intermolecular potential with 3 points) model, we generate a 12bp (12-base-pair) DNA sequence (5′-GGGCCCGGGCCC-3′) surrounded with 6177 water molecules and 22 Na^+ ions. This DNA system is impossible to model with a good DFT model. To reduce the number of atoms to a more tractable one, we only keep the water molecules and Na^+ ions within the certain range of DNA molecules, see Table 6-1. For example, within 4Å of DNA molecules, there are 192 water molecules and 11 Na^+ ions, which leads to the DNA system having a net charge of -11. That is approximately 16 explicit water molecules per DNA base pair.

Table 6-1. Number of water molecules and Na^+ ions near DNA molecules

Distance around DNA (Å)	Number of water molecules	Number of Na^+ ions	Charge
1	0	0	-22
2	0	0	-22
3	94	8	-14
4	192	11	-11
5	359	18	-4
6	554	21	-1
7	762	22	0
8	992	22	0
9	1256	22	0
10	1529	22	0

In order to reduce the computational complexity, we start our calculations with shorter DNA sequences: one 2bp DNA (5′-GG-3′), one 4bp DNA (5′-CCGG-3′), and one 6bp DNA (5′-CCCGGG-3′). The structures are obtained by trimming the above discussed 12bp DNA obtained from MD simulations. In these simpler models, the solvent molecules within 4Å of DNA and the closest Na^+ ions are retained, making the net charge zero. For 2bp case, there are 30 water molecules and 2 Na^+ ions, see Figure 6-1 (a). For 4bp case, there are 53 water molecules and 6 Na^+ ions, see Figure 6-4 (a). For 6bp case, there are 86 water molecules and 10 Na^+ ions, see Figure 6-5 (a). Note that since there are not enough Na^+ ions within 4Å of 6bp DNA molecules, we include extra Na^+ ions within 5Å to make the system charge neutral. On average, there are 14 water molecules per DNA base pair, which is comparable to previous studies [34], [144], [145]. To determine the effect of explicit water molecules along with PCM and Na^+ ions, we perform

two comparison calculations: (1) *No PCM* case, where explicit water molecules and Na^+ ions are included without turning on the PCM model. The relative dielectric constant here is $\varepsilon = 1$. (2) *With PCM* case, where apart from including the explicit water molecules and Na^+ ions, the PCM model is turned on. The relative dielectric constant in these calculations is $\varepsilon = 78.3553$. We perform similar DFT, wavefunction, and transmission calculations as in previous chapters.

6.2 Results and Discussion

For 2bp DNA (5′-GG-3′), a large difference in bandgap value between the "No PCM" (0.8422 eV) and "With PCM" (4.5829 eV) case is observed as shown in Figure 6-1 (b) and (c). Further, in the energy range from -4 to -1 eV, the "No PCM" case has six more energy levels than the "With PCM" case.

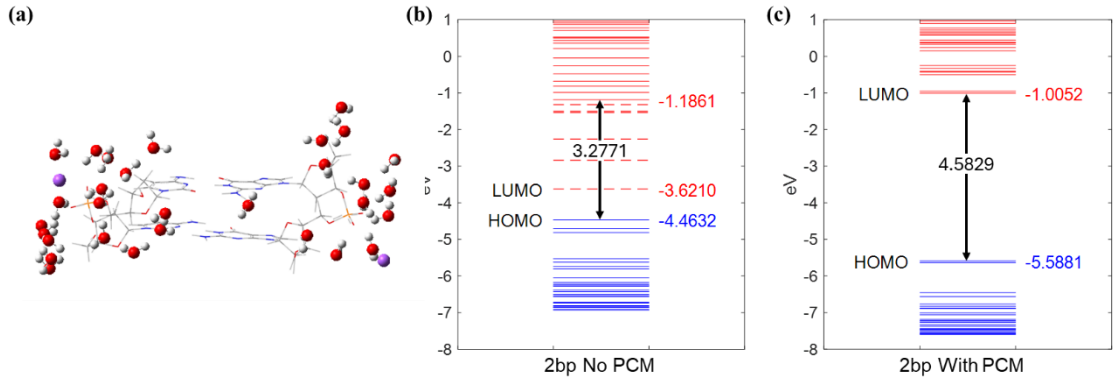


Figure 6-1. A 2-base-pair DNA model with explicit water molecules. (a) Atomic location of 30 water molecules and 2 Na^+ ions, within 4Å of DNA molecules. (b) and (c) plot the HOMO (blue) and LUMO (red) energy levels. Dash lines represent energy levels localized on water molecules or Na^+ ions. (b) "No PCM" case has a bandgap of 0.8422 eV. (c) "With PCM" case has a bandgap of 4.5829 eV.

We study the region of wavefunction localization corresponding to the various energy levels ranging from HOMO-25 to LUMO+25. For the "No PCM" case (see Figure 6-2 (c)), the wavefunctions of LUMO to LUMO+5 energy levels are localized on Na^+ ions and water molecules. As shown in Figure 6-2 (d), the lowest unoccupied level with wavefunction localization on the DNA molecule is LUMO+6 (followed by LUMO+8), which are on cytosine bases. In contrast, for the "With PCM" case, the wavefunctions are localized on the cytosine base of the DNA molecules for both LUMO and LUMO+1 energy levels. Then the wavefunctions of LUMO+2, LUMO+3, LUMO+5 and other higher energy levels are localized on the Na^+ ions and water molecules. At the HOMO and nearby energy levels, for both "No PCM" and "With PCM" cases, the wavefunction is localized on the guanine base, see Figure 6-2 (a) and Figure 6-3 (a). We also observe from Figure 6-2 (b) that in the "No PCM" case, the wavefunctions of HOMO-3 to HOMO-6 energy levels are localized on water molecules. Similarly, in the "With PCM" case, the wavefunctions of HOMO-4, HOMO-6, and HOMO-7 are all localized on water molecules,

see Figure 6-3 (c). The effect of explicit water molecules is most obvious in these occupied energy levels near HOMO.

In Figure 6-1 (b), we used dash lines to represent the energy levels whose wavefunction is localized on water molecules or Na^+ ions, until the first energy level that is localized on the DNA molecule. While the bandgap is 0.8422 eV in the "No PCM" case, the energy difference between the lowest unoccupied energy level and the highest occupied energy level whose wavefunctions are both on the DNA is 3.2771 eV. In comparison, the "With PCM" case, has HOMO and LUMO levels both localized on the DNA molecules, and the bandgap is 4.5829 eV.

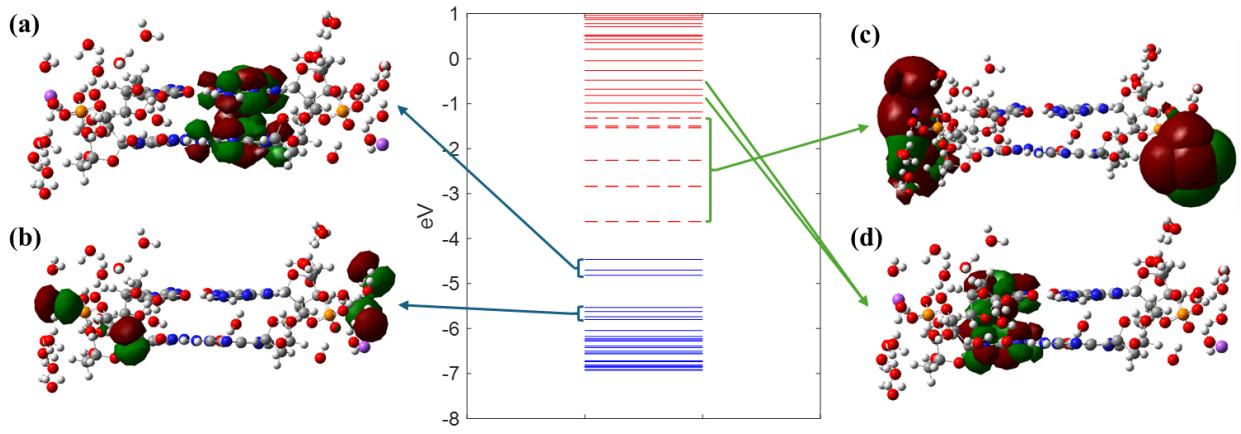


Figure 6-2. The wavefunctions of No PCM case. (a) The wavefunctions of HOMO, HOMO-1, HOMO-2 are localized on guanine bases. (b) The wavefunctions of HOMO-3 ~ HOMO-6 are localized on water molecules. (c) The wavefunctions of LUMO ~ LUMO+5 are localized on Na^+ ions and water molecules. (d) The wavefunctions of LUMO+6, LUMO+8 are localized on cytosine bases.

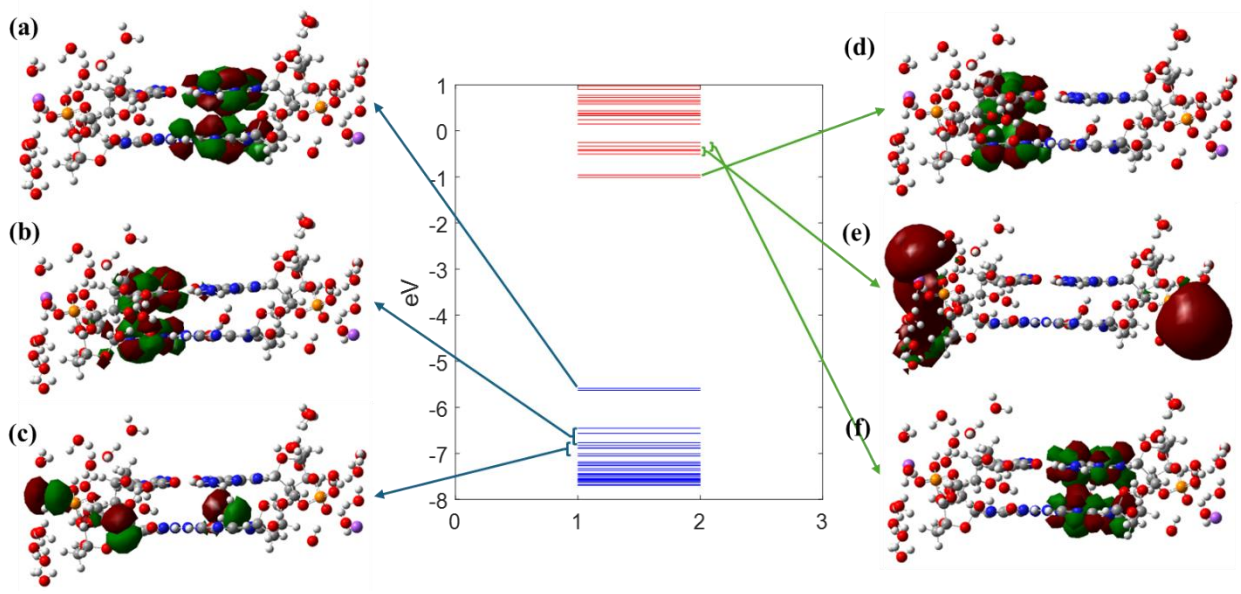


Figure 6-3. The wavefunctions of With PCM case. (a) The wavefunctions of HOMO, HOMO-1 are localized on guanine bases. (b) The wavefunctions of HOMO-2, HOMO-3, HOMO-5 are localized on cytosine bases. (c) The wavefunctions of HOMO-4, HOMO-6, HOMO-7 are localized on water molecules. (d) The wavefunctions of LUMO, LUMO+1 are localized on cytosine bases. (e) The wavefunctions of LUMO+2. LUMO+3, LUMO+5 are localized on Na^+ ions and water molecules. (f) The wavefunctions of LUMO+4, LUMO+6 are localized on guanine bases.

For 4bp DNA (5′-CCGG-3′), there is a large difference in bandgap values between the "No PCM" (1.0925 eV) and "With PCM" (4.4390 eV) cases, as seen in Figure 6-4 (b) and (c). We find that the wavefunctions of LUMO to LUMO+7 energy levels are localized on Na^+ ions and water molecules in the "No PCM" case (similar to the 2bp results). The lowest energy level for which the wavefunction is localized on the DNA molecule is LUMO+8 (followed by LUMO+10), at the cytosine and guanine bases. For "With PCM" case, while the LUMO, LUMO+1, LUMO+2, LUMO+4 wavefunctions are localized on cytosine bases, the LUMO+5, LUMO+6, LUMO+7, LUMO+10 wavefunctions are localized on the guanine bases. They each form a clear channel to enable charge transport. The wavefunctions of LUMO+3, LUMO+8, LUMO+9 and other higher energy levels are on the Na^+ ions and water molecules.

For HOMO energy levels, the "No PCM" case shows some critical changes from the 2bp case. The wavefunctions of HOMO to HOMO-4 are localized on water molecules (no longer on the bases). The wavefunctions corresponding to HOMO-5, HOMO-8, HOMO-10, HOMO-18 and HOMO-24, are localized on the guanine bases. Most other wavefunctions corresponding to the HOMO to HOMO-25 energy levels are localized on water molecules. In contrast, for the "With PCM" case, the wavefunctions of HOMO to HOMO-3 are localized on the guanine bases while the HOMO-4 to HOMO-7 wavefunctions are localized on cytosine bases. They each form

a clear channel to enable charge transport. In the lower energy levels, HOMO-10 and HOMO-11, the wavefunctions have a larger component on the explicit water molecules.

In Figure 6-4 (b), the dashed lines represent the energy levels that are localized on water molecules or Na^+ ions, until the first energy level that is localized on DNA molecules which is shown by a solid line. In the "No PCM" case, instead of a bandgap of 1.0925 eV, the difference in energy levels between the lowest unoccupied and highest occupied states that lie on the DNA molecules is 3.8684 eV. For "With PCM" case, the HOMO and LUMO levels are localized on the DNA molecules and the bandgap is 4.4390 eV. Compared with 2bp DNA, the difference of bandgap (lie on the DNA molecules) becomes smaller for 4bp DNA.

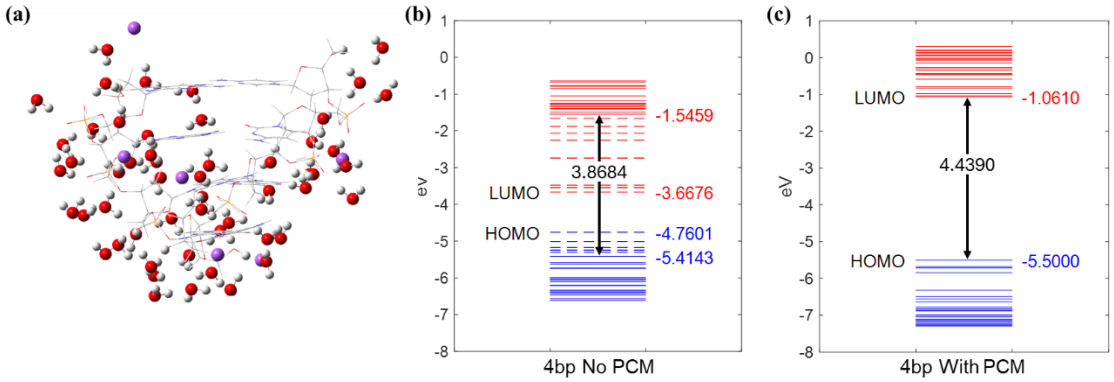


Figure 6-4. A 4-base-pair DNA model with explicit water molecules. (a) Atomic location of 53 water molecules and 6 Na^+ ions, within 4Å of DNA molecules. *(b) and (c) plot the HOMO (blue) and LUMO (red) energy levels. Dash lines represent energy levels localized on water molecules or Na^+ ions.* (b) "No PCM" case has a bandgap of 1.0925 eV. (c) "With PCM" case has a bandgap of 4.4390 eV.

For 6bp DNA (5′-CCCGGG-3′), we again observe a large difference in the bandgap for the "No PCM" (0.1711 eV) and "With PCM" (4.3557 eV) cases, see Figure 6-5 (b) and (c). The wavefunctions localization of the LUMO and HOMO energy levels is similar to the 4bp case. For the "No PCM" case, the wavefunctions of LUMO toLUMO+8 energy levels are localized on Na^+ ions and water molecules. This localization creates a path for charge transport at these energies via the Na^+ ions and water molecules. The lowest unoccupied wavefunctions with significant localization on the DNA molecule are LUMO+9 and LUMO+10, on cytosine and guanine bases. In the "With PCM" case, the wavefunctions are localized on the DNA molecules. The wavefunctions of several LUMO levels again form channels through cytosine and guanine bases to enable charge transport. The LUMO+9, LUMO+12 and other higher energy levels have wavefunctions on the Na^+ ions and water molecules.

Like the 4bp case, the wavefunctions of HOMO to HOMO-4 are localized on water molecules. In the absence of Na^+ ions, the wavefunctions on the water molecules do not form a clear channel. The wavefunctions don't localize on DNA molecules until HOMO-5, HOMO-6, HOMO-9, HOMO-12 and HOMO-21, and they are all on guanine bases. For the "With PCM"

case, the wavefunction of HOMO is localized on the guanine bases. The wavefunctions of several HOMO levels again form two channels through cytosine and guanine bases to enable charge transport. For lower lying occupied energy levels, HOMO-13 and HOMO-15, the wavefunction component on the explicit water molecules becomes more significant.

If we compute the energy difference between the lowest unoccupied (LUMO+9) and highest occupied (HOMO+5) states where the wavefunctions are on the DNA bases, we find a gap of 3.0678 eV. In Figure 6-5 (b), we used dash lines to represent the energy levels that are localized on water molecules or Na^+ ions, until the first energy level that is localized on DNA molecules is observed. For "With PCM" case, since HOMO and LUMO levels are already localized on DNA molecules, the bandgap is 4.3557 eV.

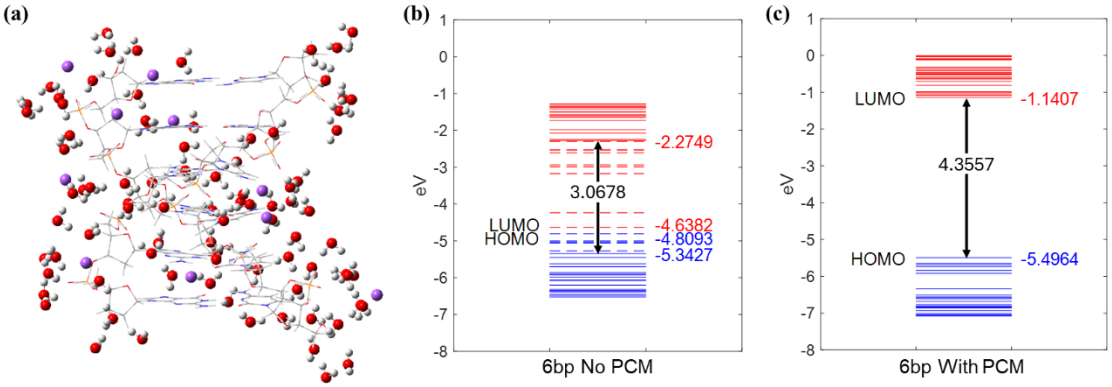


Figure 6-5. A 6-base-pair DNA model with explicit water molecules. (a) Atomic location of 86 water molecules and 10 Na^+ ions, within 5Å of DNA molecules. *(b) and (c) plot the HOMO (blue) and LUMO (red) energy levels. Dash lines represent energy levels localized on water molecules or Na^+ ions.* (b) "No PCM" case has a bandgap of 0.1711 eV. (c) "With PCM" case has a bandgap of 4.3557 eV.

After analyzing the results of adding explicit water molecules to the DNA system, we can compare these results with the ones of previous chapters, where only implicit water solvent is used for calculations. We find that our previous observations hold, but can be made more comprehensive:

- When including implicit water solvent for the DNA system ("With PCM" case/Water case).
  - The wavefunction of HOMO band is localized on Guanine or Adenine bases.
  - The wavefunction of LUMO band is localized on Cytosine or Thymine bases.
  - Additional explicit water molecules have no major influence on close-to-bandgap energy levels. The effect of explicit water molecules only starts to appear in lower occupied and higher unoccupied energy levels.
- When excluding implicit water solvent for DNA system ("No PCM" case/Dry case).
  - The wavefunction of the HOMO band is localized on Guanine or Adenine bases. But if explicit water molecules are present, the localization moves to water molecules.

- The wavefunction of LUMO level is localized on Na^+ ions and explicit water molecules (if present).
- Additional explicit water molecules help reduce the bandgap, which is similar to the effect of Na^+ ions. However, the wavefunctions of water molecules are unable to form a clear channel to enable charge transport, without involving Na^+ ions.

6.3 Summary

In Table 6-2, we summarized the HOMO, LUMO and bandgap energy values of 2bp, 4bp, and 6bp DNA. For all "No PCM" cases, the bandgap of entire system is minimal. But once we move the occupied energy levels from water to DNA molecules and the unoccupied energy levels from Na^+ ions to DNA molecules, the bandgap is greatly increased. The difference of energy levels between "No PCM" and "With PCM" cases also become much similar to each other, if we only focus on DNA molecules.

Table 6-2. The HOMO, LUMO and bandgap energy values of "No PCM" and "With PCM" cases.

Case	HOMO		LUMO		Bandgap	
	Entire System	DNA Molecule Only	Entire System	DNA Molecule Only	Entire System	DNA Molecule Only
2bp No PCM	-4.4632 eV	-4.4632 eV	-3.6210 eV	-1.1861 eV	0.8422 eV	3.2771 eV
2bp With PCM	-5.5881 eV		-1.0052 eV		4.5829 eV	
4bp No PCM	-4.7601 eV	-5.4143 eV	-3.6676 eV	-1.5459 eV	1.0925 eV	3.8684 eV
4bp With PCM	-5.5000 eV		-1.0610 eV		4.4390 eV	
6bp No PCM	-4.8093 eV	-5.3427 eV	-4.6382 eV	-2.2749 eV	0.1711 eV	3.0678 eV
6bp With PCM	-5.4964 eV		-1.1407 eV		4.3557 eV	

7 Summary and Future Work

7.1 Summary

In this thesis, we investigated the DNA electronic properties through data-driven approaches (statistical feature extraction and machine learning algorithms) and physics-driven approaches (density functional theory and Green's function method). The topics covered in this thesis are summarized below.

In Chapter 1, we discussed the fundamentals of DNA structure and the evolution of DNA sequencing techniques. We reviewed previous work, modeling challenges, experimental limitations, and potential applications of DNA electronic properties (conductance). Additionally, we emphasized the discrepancies between modeling and experimental results, paving the way for development of domain-specific machine learning (ML) algorithms to overcome these challenges.

In Chapter 2, we demonstrated the initial design of a ML based classification system for DNA sequence identification using conductance measurements from a Single Molecule Break Junction experiment. We presented a detailed description of the methods for data acquisition, pre-processing, visualization, and the XGBoost classifier. By systematically analyzing classifier performance under various parameters, we showed that over 99.5% accuracy can be achieved for short-strand DNA molecules that are structurally different with only 20~30 conductance measurements. However, the classifier model may not provide comparably high accuracy for genetic samples that differ by a single base mismatch, especially if the sample size is kept reasonably low in the interest of rapid detection and classification in practice.

In Chapter 3, we extended the target of the ML classification system to COVID-19 datasets, which primarily have single base mismatches. Since the performance of the algorithm discussed in Chapter 2 was inadequate, we experimented with both 1D and 2D conductance probability distributions. The 2D distributions depend on conductance as a function of inter-electrode distance. We studied the accuracy of detection with and without averaging of the conductance distribution over experimental parameters. In conjunction with 2D distributions, we have experimented with a CNN paired with an XGBoost classifier. We showed that (i) averaged conductance distributions have a significant impact on classifier accuracy, (ii) 2D conductance distributions are beneficial for some sequences, and (iii) the backend XGBoost classifier whose input is a feature vector extracted from a CNN classifier provides better accuracy than a densely connected classifier usually adopted with a CNN.

In Chapter 4, we addressed the low signal-to-noise ratio of SMBJ measurements by proposing a Piecewise Linear Approximation (PLA) method. By utilizing the PLA method, conductance plateau segments are effectively and precisely isolated, enhancing the clarity of conductance data, mitigating the noise inherent in SMBJ measurements, and increasing average accuracy up to 4.22%. The resulting plateau-only conductance histograms enabled more

effective training of machine learning models, resulting in substantial improvements in classification accuracy across various configurations, which increased from 83.96% (initial design) to 96.31% (best configuration). Especially with a smaller sample size, as few as 5, our classification system is able to achieve a high accuracy of over 80% for most variants (excluding Beta_MM1 and Beta_MM3 variants, which turned out to be particularly hard to classify). Impressively, using a sample size of 30 sample, the classification accuracy of Beta_MM1 and Beta_MM3 variants increases from about 83% to 97% by tuning the cutoff slope, a tunable parameter in the PLA method. These findings highlight the potential of PLA as an effective preprocessing tool to enhance the quality of SMBJ data. Additionally, the robustness of this method across multiple machine learning configurations indicates its versatility, suggesting that PLA could be employed in a broader range of single-molecule analysis techniques, including other types of experimental setups involving molecular junctions or time-series data.

In Chapter 5, we investigated the role of solvent dielectric constant (implicit water molecules) and Na^+ counterions on the conductance of B-DNA using DFT (Density Functional Theory) and charge transport calculations (Green's Function and Büttiker Probes). We presented a detailed description of the methods for DFT and charge transport calculations. Our simulation results demonstrate that it is essential to consider counterions separately when analyzing the DNA conductance experiments done in a dry environment, but not necessarily with a water solvent. The higher the dielectric constant, the higher the charge screening effect, thus lowering the coupling between Na^+ ions and DNA molecules. By performing calculations on six different DNA sequences and four different basis sets of DFT calculations, we demonstrated the generalizability of our results.

In Chapter 6, we explored the influence of explicit water molecules on the conductance of B-DNA using the same modeling approach as in the previous chapter. We find that our observations from implicit water molecules still hold valid. For water solvent conditions, additional explicit water molecules have no major influence on close-to-bandgap energy levels. The effect of explicit water molecules only starts to appear in lower occupied energy levels (less than HOMO) and higher unoccupied energy levels (larger than LUMO). For dry conditions, additional explicit water molecules help reduce the bandgap, which is similar to the effect of Na^+ ions. However, the wavefunctions of water molecules are unable to form a clear channel to enable charge transport, without involving Na^+ ions.

7.2 Future Work

While this study has advanced the understanding of DNA electronic properties and sequence identification, several directions remain open for further exploration. First, we could optimize the PLA for various molecular targets and expand its application to different types of single-molecule systems, such as protein interactions, chemical sensing, and real-time biological processes. Designing a highly accurate classifier model operating on histograms constructed from a single trace is most beneficial. We expect that generative AI models would be beneficial

for achieving that gold standard. Adoption of explainable AI and feature importance ranking procedures can help in providing a strong connection between the experimental and modeling results.

Building on the expansive knowledge earned through the data-driven approach, theory development becomes even more critical. It is crucial to move beyond traditional physics-based theories and modeling approaches which are ineffective in a highly stochastic environment as in the SMBJ approach. The majority of existing theoretical simulation results have significant discrepancies from experimental ones, which can be as high as two orders of magnitude when predicting the conductance of DNA molecules. Since it is nearly impossible to model a DNA system in a realistic manner when considering all possible factors (such as DNA internal structure, environmental fluctuations, DNA-contact configuration), experiment-based theories that leverage extensive experimental data would be a beneficial direction. For instance, it may be possible to develop a simplified DNA dynamic system or a coarse-grained DNA structure to mimic an SMBJ experiment.

8 References

[1] K. Wang, "DNA-based single-molecule electronics: From concept to function," *J Funct Biomater*, vol. 9, no. 1, p. 8, 2018, doi: 10.3390/jfb9010008.

[2] V. Bhalla, R. P. Bajpai, and L. M. Bharadwaj, "DNA electronics," *EMBO Rep*, vol. 4, no. 5, pp. 442–445, May 2003, doi: 10.1038/sj.embor.embor834.

[3] H. Cohen, C. Nogues, R. Naaman, and D. Porath, "Direct measurement of electrical transport through single DNA molecules of complex sequence," *Proceedings of the National Academy of Sciences*, vol. 102, no. 33, pp. 11589–11593, Aug. 2005, doi: 10.1073/pnas.0505272102.

[4] H.-W. Fink and C. Schönenberger, "Electrical conduction through DNA molecules," *Nature*, vol. 398, no. 6726, pp. 407–410, Apr. 1999, doi: 10.1038/18855.

[5] A. Y. Kasumov *et al.*, "Proximity-Induced Superconductivity in DNA," *Science (1979)*, vol. 291, no. 5502, pp. 280–282, Jan. 2001, doi: 10.1126/science.291.5502.280.

[6] K.-H. Yoo *et al.*, "Electrical Conduction through Poly(dA)-Poly(dT) and Poly(dG)-Poly(dC) DNA Molecules," *Phys Rev Lett*, vol. 87, no. 19, p. 198102, Oct. 2001, doi: 10.1103/PhysRevLett.87.198102.

[7] A. J. Storm, J. van Noort, S. de Vries, and C. Dekker, "Insulating behavior for DNA molecules between nanoelectrodes at the 100 nm length scale," *Appl Phys Lett*, vol. 79, no. 23, pp. 3881–3883, Dec. 2001, doi: 10.1063/1.1421086.

[8] Y. Zhang, R. H. Austin, J. Kraeft, E. C. Cox, and N. P. Ong, "Insulating Behavior of λ-DNA on the Micron Scale," *Phys Rev Lett*, vol. 89, no. 19, p. 198102, Oct. 2002, doi: 10.1103/PhysRevLett.89.198102.

[9] Xu, Zhang, Li, and 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.

[10] J. Hihath, B. Xu, P. Zhang, and N. Tao, "Study of single-nucleotide polymorphisms by means of electrical conductance measurements," *Proceedings of the National Academy of Sciences*, vol. 102, no. 47, pp. 16979–16983, Nov. 2005, doi: 10.1073/pnas.0505175102.

[11] C. Bruot, L. Xiang, J. L. Palma, and N. Tao, "Effect of Mechanical Stretching on DNA Conductance," *ACS Nano*, vol. 9, no. 1, pp. 88–94, Jan. 2015, doi: 10.1021/nn506280t.

[12] K. Wang *et al.*, "Structure determined charge transport in single DNA molecule break junctions," *Chem. Sci.*, vol. 5, no. 9, pp. 3425–3431, 2014, doi: 10.1039/C4SC00888J.

[13] B. Xu, J. Hamill, and K. Wang, “Characterizing molecular junctions through the mechanically controlled break-junction approach,” *Reports in Electrochemistry*, vol. 4, pp. 1–11, May 2014, doi: 10.2147/RIE.S46629.

[14] K. Wang and B. Xu, “Modulation and Control of Charge Transport Through Single-Molecule Junctions,” *Top Curr Chem*, vol. 375, no. 1, p. 17, Feb. 2017, doi: 10.1007/s41061-017-0105-z.

[15] J. M. Artés, Y. Li, J. Qi, M. P. Anantram, and J. Hihath, “Conformational gating of DNA conductance,” *Nat Commun*, vol. 6, no. 1, p. 8870, Dec. 2015, doi: 10.1038/ncomms9870.

[16] 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.

[17] S. Afsari, L. E. Korshoj, G. R. Abel, S. Khan, A. Chatterjee, and P. Nagpal, “Quantum Point Contact Single-Nucleotide Conductance for DNA and RNA Sequence Identification,” *ACS Nano*, vol. 11, no. 11, pp. 11169–11181, Nov. 2017, doi: 10.1021/acsnano.7b05500.

[18] G. R. Abel, L. E. Korshoj, P. B. Otoupal, S. Khan, A. Chatterjee, and P. Nagpal, “Nucleotide and structural label identification in single RNA molecules with quantum tunneling spectroscopy,” *Chem Sci*, vol. 10, no. 4, pp. 1052–1063, 2019, doi: 10.1039/C8SC03354D.

[19] M. H. F. Wilkins, A. R. Stokes, and H. R. Wilson, “Molecular structure of nucleic acids: Molecular structure of deoxypentose nucleic acids,” *Nature*, vol. 171, no. 4356, pp. 738–740, 1953, doi: 10.1038/171738a0.

[20] K. A. Wetterstrand, “DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).” Accessed: Aug. 09, 2020. [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

[21] J. Ritchie, “Probabilistic DNA evidence: The laypersons interpretation,” *Australian Journal of Forensic Sciences*, vol. 47, no. 4, pp. 440–449, Oct. 2015, doi: 10.1080/00450618.2014.992472.

[22] 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.

[23] A. Grada and K. Weinbrecht, “Next-Generation Sequencing: Methodology and Application,” *Journal of Investigative Dermatology*, vol. 133, no. 8, pp. 1–4, Aug. 2013, doi: 10.1038/jid.2013.248.

[24] M. Kircher and J. Kelso, “High-throughput DNA sequencing - Concepts and limitations,” Jun. 01, 2010, *John Wiley & Sons, Ltd*. doi: 10.1002/bies.200900181.

[25] S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie, and Q. Gouil, “Opportunities and challenges in long-read sequencing data analysis,” Feb. 07, 2020, *BioMed Central Ltd.* doi: 10.1186/s13059-020-1935-5.

[26] M. O. Pollard, D. Gurdasani, A. J. Mentzer, T. Porter, and M. S. Sandhu, “Long reads: their purpose and place,” Aug. 01, 2018, *NLM (Medline)*. doi: 10.1093/hmg/ddy177.

[27] P. B. Woiczikowski, T. Kuba, R. Gutírrez, R. A. Caetano, G. Cuniberti, and M. Elstner, “Combined density functional theory and Landauer approach for hole transfer in DNA along classical molecular dynamics trajectories,” *Journal of Chemical Physics*, vol. 130, no. 21, p. 215104, 2009, doi: 10.1063/1.3146905.

[28] L. E. Korshoj, S. Afsari, S. Khan, A. Chatterjee, and P. Nagpal, “Single Nucleobase Identification Using Biophysical Signatures from Nanoelectronic Quantum Tunneling,” *Small*, vol. 13, no. 11, pp. 1–10, 2017, doi: 10.1002/smll.201603033.

[29] D. Dulić, S. Tuukkanen, C.-L. Chung, A. Isambert, P. Lavie, and A. Filoramo, “Direct conductance measurements of short single DNA molecules in dry conditions,” *Nanotechnology*, vol. 20, no. 11, p. 115502, Mar. 2009, doi: 10.1088/0957-4484/20/11/115502.

[30] A. K. Mahapatro, D. B. Janes, K. J. Jeong, and G. U. Lee, “Electrical Behavior of Nano-scale Junctions with Well Engineered Double Stranded DNA Molecules,” in *2006 Sixth IEEE Conference on Nanotechnology*, IEEE, 2006, pp. 66–69. doi: 10.1109/NANO.2006.247568.

[31] A. K. Mahapatro, G. U. Lee, K. J. Jeong, and D. B. Janes, “Stable and reproducible electronic conduction through DNA molecular junctions,” *Appl Phys Lett*, vol. 95, no. 8, p. 083106, Aug. 2009, doi: 10.1063/1.3186056.

[32] A. K. Mahapatro, G. U. Lee, K. J. Jeong, and D. B. Janes, “Stable and reproducible electronic conduction through DNA molecular junctions,” *Appl Phys Lett*, vol. 95, no. 8, p. 083106, Aug. 2009, doi: 10.1063/1.3186056.

[33] T. Kubař, R. Gutiérrez, U. Kleinekathöfer, G. Cuniberti, and M. Elstner, “Modeling charge transport in DNA using multi-scale methods,” *physica status solidi (b)*, vol. 250, no. 11, pp. 2277–2287, Nov. 2013, doi: 10.1002/pssb.201349148.

[34] M. Wolter, M. Elstner, and T. Kubař, “Charge transport in desolvated DNA,” *J Chem Phys*, vol. 139, no. 12, p. 125102, Sep. 2013, doi: 10.1063/1.4821594.

[35] M. Feig and B. M. Pettitt, “A molecular simulation picture of DNA hydration around A- and B-DNA,” *Biopolymers*, vol. 48, no. 4, p. 199, 1998, doi: 10.1002/(SICI)1097-0282(1998)48:4<199::AID-BIP2>3.0.CO;2-5.

[36] T. Kubař and M. Elstner, “What governs the charge transfer in DNA? The role of DNA conformation and environment,” *Journal of Physical Chemistry B*, vol. 112, no. 29, pp. 8788–8798, 2008, doi: 10.1021/jp803661f.

[37] 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.

[38] 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 Genomics*, vol. 22, no. 1, p. 525, Dec. 2021, doi: 10.1186/s12864-021-07841-6.

[39] J. Berashevich and T. Chakraborty, “Water induced weakly bound electrons in DNA,” *J Chem Phys*, vol. 128, no. 23, p. 235101, Jun. 2008, doi: 10.1063/1.2939248.

[40] R. Gutiérrez, R. Caetano, P. B. Woiczikowski, T. Kubar, M. Elstner, and G. Cuniberti, “Structural fluctuations and quantum transport through DNA molecular wires: a combined molecular dynamics and model Hamiltonian approach,” *New J Phys*, vol. 12, no. 2, p. 023022, Feb. 2010, doi: 10.1088/1367-2630/12/2/023022.

[41] R. Venkatramani, S. Keinan, A. Balaeff, and D. N. Beratan, “Nucleic acid charge transfer: Black, white and gray,” *Coord Chem Rev*, vol. 255, no. 7–8, pp. 635–648, Apr. 2011, doi: 10.1016/j.ccr.2010.12.010.

[42] E. Wierzbinski *et al.*, “Charge Transfer through Modified Peptide Nucleic Acids,” *Langmuir*, vol. 28, no. 4, pp. 1971–1981, Jan. 2012, doi: 10.1021/la204445u.

[43] M. Wolter, M. Elstner, U. Kleinekathöfer, and T. Kubař, “Microsecond Simulation of Electron Transfer in DNA: Bottom-Up Parametrization of an Efficient Electron Transfer Model Based on Atomistic Details,” *J Phys Chem B*, vol. 121, no. 3, pp. 529–549, Jan. 2017, doi: 10.1021/acs.jpcb.6b11384.

[44] Y. Zhang *et al.*, “Conformationally Gated Charge Transfer in DNA Three-Way Junctions,” *J Phys Chem Lett*, vol. 6, no. 13, pp. 2434–2438, Jul. 2015, doi: 10.1021/acs.jpclett.5b00863.

[45] Ch. Adessi, S. Walch, and M. P. Anantram, “Environment and structure influence on DNA conduction,” *Phys Rev B*, vol. 67, no. 8, p. 081405, Feb. 2003, doi: 10.1103/PhysRevB.67.081405.

[46] M. Kulkarni and A. Mukherjee, “Understanding B-DNA to A-DNA transition in the right-handed DNA helix: Perspective from a local to global transition,” *Prog Biophys Mol Biol*, vol. 128, pp. 63–73, Sep. 2017, doi: 10.1016/j.pbiomolbio.2017.05.009.

[47] A. Ghosh and M. Bansal, “A glossary of DNA structures from A to Z,” *Acta Crystallogr D Biol Crystallogr*, vol. 59, no. 4, pp. 620–626, Apr. 2003, doi: 10.1107/S0907444903003251.

[48] P. K. Maiti and B. Bagchi, “Structure and dynamics of DNA-dendrimer complexation: Role of counterions, water, and base pair sequence,” *Nano Lett*, vol. 6, no. 11, pp. 2478–2485, 2006, doi: 10.1021/nl061609m.

[49] C. Adessi and M. P. Anantram, “Influence of counter-ion-induced disorder in DNA conduction,” *Appl Phys Lett*, vol. 82, no. 14, pp. 2353–2355, Apr. 2003, doi: 10.1063/1.1563811.

[50] D. M. Sagar *et al.*, “High-Throughput Block Optical DNA Sequence Identification,” *Small*, vol. 14, no. 4, pp. 1–9, 2018, doi: 10.1002/smll.201703165.

[51] J. C. Ribot, A. Chatterjee, and P. Nagpal, “Measurements of single nucleotide electronic states as nanoelectronic fingerprints for identification of DNA nucleobases, their protonated and unprotonated states, isomers, and tautomers,” *Journal of Physical Chemistry B*, vol. 119, no. 15, pp. 4968–4974, 2015, doi: 10.1021/acs.jpcb.5b01403.

[52] M. Kolmogorov, E. Kennedy, Z. Dong, G. Timp, and P. A. Pevzner, “Single-molecule protein identification by sub-nanopore sensors,” *PLoS Comput Biol*, vol. 13, no. 5, pp. 1–14, 2017, doi: 10.1371/journal.pcbi.1005356.

[53] D. Cabosart *et al.*, “A reference-free clustering method for the analysis of molecular break-junction measurements,” *Appl Phys Lett*, vol. 114, no. 14, 2019, doi: 10.1063/1.5089198.

[54] J. M. Hamill, X. T. Zhao, G. Mészáros, M. R. Bryce, and M. Arenz, “Fast Data Sorting with Modified Principal Component Analysis to Distinguish Unique Single Molecular Break Junction Trajectories,” *Phys Rev Lett*, vol. 120, no. 1, 2018, doi: 10.1103/PhysRevLett.120.016601.

[55] 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.

[56] Xu, Zhang, Li, and 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.

[57] T. Nishino and P. T. Bui, “Direct electrical single-molecule detection of DNA through electron transfer induced by hybridization,” *Chemical Communications*, vol. 49, no. 33, p. 3437, 2013, doi: 10.1039/c3cc38992h.

[58] K. G. Gunasinghe Pattiya Arachchillage *et al.*, “A single-molecule RNA electrical biosensor for COVID-19,” *Biosens Bioelectron*, vol. 239, p. 115624, Nov. 2023, doi: 10.1016/j.bios.2023.115624.

[59] 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.

[60] 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.

[61] B. Zhang *et al.*, “Role of contacts in long-range protein conductance,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 13, pp. 5886–5891, Mar. 2019, doi: 10.1073/pnas.1819674116.

[62] 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.

[63] 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,” *Journal of Physics: Condensed Matter*, vol. 24, no. 16, p. 164211, Apr. 2012, doi: 10.1088/0953-8984/24/16/164211.

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

[65] H. Li *et al.*, “Large Variations in the Single-Molecule Conductance of Cyclic and Bicyclic Silanes,” *J Am Chem Soc*, vol. 140, no. 44, pp. 15080–15088, Nov. 2018, doi: 10.1021/jacs.8b10296.

[66] 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.

[67] Z. Wang *et al.*, “Electrochemically controlled rectification in symmetric single-molecule junctions,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 39, Sep. 2022, doi: 10.1073/pnas.2122183119.

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

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

[70] 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, doi: 10.1002/asia.202200179.

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

[72] S. Tao *et al.*, “Revealing conductance variation of molecular junctions based on an unsupervised data analysis approach,” *Electrochim Acta*, vol. 449, p. 142225, May 2023, doi: 10.1016/j.electacta.2023.142225.

[73] 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.

[74] 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,” *The Journal of Physical Chemistry C*, vol. 124, no. 33, pp. 18302–18315, Aug. 2020, doi: 10.1021/acs.jpcc.0c03612.

[75] S. V. Aradhya and L. Venkataraman, “Single-molecule junctions beyond electronic transport,” *Nature Nanotechnology 2013 8:6*, vol. 8, no. 6, pp. 399–410, Jun. 2013, doi: 10.1038/nnano.2013.91.

[76] H. Chen, Y. Li, and S. Chang, “Hybrid Molecular-Junction Mapping Technique for Simultaneous Measurements of Single-Molecule Electronic Conductance and Its Corresponding Binding Geometry in a Tunneling Junction,” *Anal Chem*, vol. 92, no. 9, pp. 6423–6429, May 2020, doi: 10.1021/ACS.ANALCHEM.9B05549/ASSET/IMAGES/LARGE/AC9B05549_0003.JPEG.

[77] T. A. Su *et al.*, “Silicon ring strain creates high-conductance pathways in single-molecule circuits,” *J Am Chem Soc*, vol. 135, no. 49, pp. 18331–18334, Dec. 2013, doi: 10.1021/JA410656A/SUPPL_FILE/JA410656A_SI_003.CIF.

[78] E. Leary, A. La Rosa, M. T. González, G. Rubio-Bollinger, N. Agraït, and N. Martín, "Incorporating single molecules into electrical circuits. The role of the chemical anchoring group," *Chem Soc Rev*, vol. 44, no. 4, pp. 920–942, Feb. 2015, doi: 10.1039/C4CS00264D.

[79] C. R. Parker *et al.*, "A comprehensive study of extended tetrathiafulvalene cruciform molecules for molecular electronics: Synthesis and electrical transport measurements," *J Am Chem Soc*, vol. 136, no. 47, pp. 16497–16507, Nov. 2014, doi: 10.1021/JA509937K/SUPPL_FILE/JA509937K_SI_002.CIF.

[80] R. Frisenda, V. A. E. C. Janssen, F. C. Grozema, H. S. J. Van Der Zant, and N. Renaud, "Mechanically controlled quantum interference in individual π-stacked dimers," *Nature Chemistry 2016 8:12*, vol. 8, no. 12, pp. 1099–1104, Aug. 2016, doi: 10.1038/nchem.2588.

[81] E. Leary *et al.*, "The Role of Oligomeric Gold-Thiolate Units in Single-Molecule Junctions of Thiol-Anchored Molecules," *Journal of Physical Chemistry C*, vol. 122, no. 6, pp. 3211–3218, Feb. 2018, doi: 10.1021/ACS.JPCC.7B11104/ASSET/IMAGES/LARGE/JP-2017-11104E_0007.JPEG.

[82] D. Z. Manrique *et al.*, "A quantum circuit rule for interference effects in single-molecule electrical junctions," *Nature Communications 2015 6:1*, vol. 6, no. 1, pp. 1–8, Mar. 2015, doi: 10.1038/ncomms7389.

[83] Y. Hasegawa, T. Harashima, Y. Jono, T. Seki, M. Kiguchi, and T. Nishino, "Kinetic investigation of a chemical process in single-molecule junction," *Chemical Communications*, vol. 56, no. 2, pp. 309–312, Dec. 2019, doi: 10.1039/C9CC08383A.

[84] A. Mishchenko *et al.*, "Single-molecule junctions based on nitrile-terminated biphenyls: A promising new anchoring group," *J Am Chem Soc*, vol. 133, no. 2, pp. 184–187, Jan. 2011, doi: 10.1021/JA107340T/SUPPL_FILE/JA107340T_SI_001.PDF.

[85] Z.-H. Zhao *et al.*, "Single-Molecule Conductance through an Isoelectronic B–N Substituted Phenanthrene Junction," *J Am Chem Soc*, vol. 142, no. 18, pp. 8068–8073, May 2020, doi: 10.1021/JACS.0C00879.

[86] Xiao, Xu, and N. J. Tao, "Measurement of Single Molecule Conductance: Benzenedithiol and Benzenedimethanethiol," *Nano Lett*, vol. 4, no. 2, pp. 267–271, Feb. 2004, doi: 10.1021/nl035000m.

[87] B. Zhang *et al.*, "Observation of giant conductance fluctuations in a protein," *Nano Futures*, vol. 1, no. 3, p. 035002, Nov. 2017, doi: 10.1088/2399-1984/aa8f91.

[88] K. G. G. Pattiya Arachchillage *et al.*, “Electrical detection of RNA cancer biomarkers at the single-molecule level,” *Sci Rep*, vol. 13, no. 1, p. 12428, Aug. 2023, doi: 10.1038/s41598-023-39450-6.

[89] E. Leary *et al.*, “Detecting Mechanochemical Atropisomerization within an STM Break Junction,” *J Am Chem Soc*, vol. 140, no. 2, pp. 710–718, Jan. 2018, doi: 10.1021/jacs.7b10542.

[90] I. J. Planje *et al.*, “Selective Anchoring Groups for Molecular Electronic Junctions with ITO Electrodes,” *ACS Sens*, vol. 6, no. 2, pp. 530–537, Feb. 2021, doi: 10.1021/acssensors.0c02205.

[91] Z. Aminiranjbar *et al.*, “Identifying SARS-CoV-2 Variants Using Single-Molecule Conductance Measurements,” *ACS Sens*, May 2024, doi: 10.1021/acssensors.3c02734.

[92] Y. Wang, V. Khandelwal, A. K. Das, and M. P. Anantram, “Classification of DNA Sequences: Performance Evaluation of Multiple Machine Learning Methods,” in *2022 IEEE 22nd International Conference on Nanotechnology (NANO)*, IEEE, Jul. 2022, pp. 333–336. doi: 10.1109/NANO54668.2022.9928773.

[93] S. Ghosh and S. Maka, “A fuzzy clustering based technique for piecewise affine approximation of a class of nonlinear systems,” *Commun Nonlinear Sci Numer Simul*, vol. 15, no. 9, pp. 2235–2244, Sep. 2010, doi: 10.1016/J.CNSNS.2009.09.017.

[94] T. Lima Silva, E. Camponogara, A. Furtado Teixeira, and S. Sunjerga, “Modeling of flow splitting for production optimization in offshore gas-lifted oil fields: Simulation validation and applications,” *J Pet Sci Eng*, vol. 128, pp. 86–97, Apr. 2015, doi: 10.1016/J.PETROL.2015.02.018.

[95] D. Lemire, “A better alternative to piecewise linear time series segmentation,” *Proc West Mark Ed Assoc Conf*, pp. 545–550, 2007, doi: 10.1137/1.9781611972771.59.

[96] R. Machné, D. B. Murray, and P. F. Stadler, “Similarity-Based Segmentation of Multi-Dimensional Signals,” *Scientific Reports 2017 7:1*, vol. 7, no. 1, pp. 1–11, Sep. 2017, doi: 10.1038/s41598-017-12401-8.

[97] E. Camponogara and L. F. Nazari, “Models and Algorithms for Optimal Piecewise-Linear Function Approximation,” *Math Probl Eng*, vol. 2015, 2015, doi: 10.1155/2015/876862.

[98] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, “Cancer Genome Landscapes,” *Science (1979)*, vol. 339, no. 6127, pp. 1546–1558, Mar. 2013, doi: 10.1126/science.1235122.

[99] “WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data.” Accessed: Apr. 06, 2022. [Online]. Available: https://covid19.who.int/

[100] “COVID-19 Map - Johns Hopkins Coronavirus Resource Center.” Accessed: Apr. 06, 2022. [Online]. Available: https://coronavirus.jhu.edu/map.html

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

[102] C. J. L. Murray, “COVID-19 will continue but the end of the pandemic is near,” *The Lancet*, vol. 399, no. 10323, pp. 417–419, 2022, doi: 10.1016/S0140-6736(22)00100-3.

[103] “Overview of Testing for SARS-CoV-2, the virus that causes COVID-19 | CDC.” Accessed: Apr. 06, 2022. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/hcp/testing-overview.html

[104] C. Bracis *et al.*, “Widespread testing, case isolation and contact tracing may allow safe school reopening with continued moderate physical distancing: A modeling analysis of King County, WA data,” *Infect Dis Model*, vol. 6, pp. 24–35, 2021, doi: 10.1016/j.idm.2020.11.003.

[105] M. J. Mina and K. G. Andersen, “COVID-19 testing: One size does not fit all,” *Science (1979)*, vol. 371, no. 6525, pp. 126–127, Jan. 2021, doi: 10.1126/science.abe9187.

[106] Q. Ma *et al.*, “Global Percentage of Asymptomatic SARS-CoV-2 Infections Among the Tested Population and Individuals With Confirmed COVID-19 Diagnosis,” *JAMA Netw Open*, vol. 4, no. 12, p. e2137257, Dec. 2021, doi: 10.1001/jamanetworkopen.2021.37257.

[107] P. Sah *et al.*, “Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis,” *Proc Natl Acad Sci U S A*, vol. 118, no. 34, pp. 1–12, 2021, doi: 10.1073/pnas.2109229118.

[108] D. B. Larremore *et al.*, “Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening,” *Sci Adv*, vol. 7, no. 1, pp. 1–11, 2021, doi: 10.1126/sciadv.abd5393.

[109] A. Deckert *et al.*, “Effectiveness and cost-effectiveness of four different strategies for SARS-CoV-2 surveillance in the general population (CoV-Surv Study): study protocol for a two-factorial randomized controlled multi-arm trial with cluster sampling,” *Trials*, vol. 22, no. 1, p. 656, Dec. 2021, doi: 10.1186/s13063-021-05619-5.

[110] “Molecular Diagnostic Tests for SARS-CoV-2 | FDA.” Accessed: Apr. 08, 2022. [Online]. Available: https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-emergency-use-authorizations-medical-devices/in-vitro-diagnostics-euas-molecular-diagnostic-tests-sars-cov-2

[111] “Antigen Diagnostic Tests for SARS-CoV-2 | FDA.” Accessed: Apr. 08, 2022. [Online]. Available: https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-

emergency-use-authorizations-medical-devices/in-vitro-diagnostics-euas-antigen-diagnostic-tests-sars-cov-2

[112] R. Weissleder, H. Lee, J. Ko, and M. J. Pittet, "COVID-19 diagnostics in context," *Sci Transl Med*, vol. 12, no. 546, pp. 1–6, 2020, doi: 10.1126/scitranslmed.abc1931.

[113] B. Kayaaslan, A. Kaya Kalem, F. Eser, I. Hasanoglu, and R. Guner, "The additional contribution of second nasopharyngeal PCR to COVID-19 diagnosis in patients with negative initial test," *Infect Dis Now*, vol. 51, no. 2, pp. 194–196, Mar. 2021, doi: 10.1016/j.medmal.2020.10.003.

[114] A. Pekosz *et al.*, "Antigen-Based Testing but Not Real-Time Polymerase Chain Reaction Correlates with Severe Acute Respiratory Syndrome Coronavirus 2 Viral Culture," *Clinical Infectious Diseases*, vol. 73, no. 9, pp. E2861–E2866, 2021, doi: 10.1093/cid/ciaa1706.

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

[116] "XGBoost Python Package — xgboost 1.3.0-SNAPSHOT documentation." Accessed: Sep. 18, 2020. [Online]. Available: https://xgboost.readthedocs.io/en/latest/python/index.html

[117] L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," 2008.

[118] M. Wattenberg, F. Viégas, and I. Johnson, "How to Use t-SNE Effectively," *Distill*, vol. 1, no. 10, Oct. 2016, doi: 10.23915/distill.00002.

[119] D. Kobak and P. Berens, "The art of using t-SNE for single-cell transcriptomics," *Nat Commun*, vol. 10, no. 1, pp. 1–14, Dec. 2019, doi: 10.1038/s41467-019-13056-x.

[120] T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. in Springer Series in Statistics. New York, NY: Springer New York, 2001. doi: 10.1007/978-0-387-21606-5.

[121] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[122] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[123] U. Von Luxburg, “A tutorial on spectral clustering,” *Stat Comput*, vol. 17, no. 4, pp. 395–416, 2007, doi: 10.1007/s11222-007-9033-z.

[124] A. Y. Ng, A. Y. Ng, M. I. Jordan, and Y. Weiss, “On Spectral Clustering: Analysis and an algorithm,” *Adv Neural Inf Process Syst*, vol. 14, pp. 849--856, 2001.

[125] J. He, O. Sankey, M. Lee, N. Tao, X. Li, and S. Lindsay, “Measuring single molecule conductance with break junctions,” *Faraday Discuss*, vol. 131, no. 0, pp. 145–154, Jan. 2006, doi: 10.1039/B508434M.

[126] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.6980

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

[128] “scipy.spatial.distance.jensenshannon — SciPy v1.12.0 Manual.” Accessed: Jan. 31, 2024. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.jensenshannon.html

[129] Y. Wang, H. Wang, A. K. Das, and M. P. Anantram, “A ML Framework for Genetic Sequence Identification using 2D Electrical Conductance Probability Distributions from Mixed Data Sets,” *IEEE Transactions on Computational Biology and Bioinformatics*, pp. 1–11, 2025, doi: 10.1109/TCBBIO.2025.3536282.

[130] Y. Li *et al.*, “Comparing Charge Transport in Oligonucleotides: RNA:DNA Hybrids and DNA Duplexes,” *J Phys Chem Lett*, vol. 7, no. 10, pp. 1888–1894, May 2016, doi: 10.1021/acs.jpclett.6b00749.

[131] T. J. Macke and D. A. Case, “Molecular Modeling of Nucleic Acids,” edited by N. B. Leontes and J. SantaLucia Jr., Ed., American Chemical Society, Washington, DC, 1998, pp. 379–393.

[132] J. Qi, N. Edirisinghe, M. G. Rabbani, and M. P. Anantram, “Unified model for conductance through DNA with the Landauer-Büttiker formalism,” *Phys Rev B*, vol. 87, no. 8, p. 085404, Feb. 2013, doi: 10.1103/PhysRevB.87.085404.

[133] T. Yamamoto, T. Uda, T. Yamasaki, and T. Ohno, “Hydration effect on the optical property of a DNA fiber: First-principles and molecular dynamics studies,” *Physical Chemistry Chemical Physics*, vol. 12, no. 32, p. 9300, 2010, doi: 10.1039/b924678a.

[134] P. Slavíček, B. Winter, M. Faubel, S. E. Bradforth, and P. Jungwirth, “Ionization Energies of Aqueous Nucleic Acids: Photoelectron Spectroscopy of Pyrimidine Nucleosides and ab

Initio Calculations," *J Am Chem Soc*, vol. 131, no. 18, pp. 6460–6467, May 2009, doi: 10.1021/ja8091246.

[135] E. Pluhařová, P. Slavíček, and P. Jungwirth, "Modeling Photoionization of Aqueous DNA and Its Components," *Acc Chem Res*, vol. 48, no. 5, pp. 1209–1217, May 2015, doi: 10.1021/ar500366z.

[136] M. J. Frisch *et al.*, "Gaussian 16, Revision C.01," 2016, *Gaussian, Inc., Wallingford CT*.

[137] H. Mohammad *et al.*, "Role of intercalation in the electrical properties of nucleic acids for use in molecular electronics," *Nanoscale Horiz*, vol. 6, no. 8, pp. 651–660, 2021, doi: 10.1039/D1NH00211B.

[138] C. J. O. Verzijl, J. S. Seldenthuis, and J. M. Thijssen, "Applicability of the wide-band limit in DFT-based molecular transport calculations," *J Chem Phys*, vol. 138, no. 9, p. 094102, Mar. 2013, doi: 10.1063/1.4793259.

[139] H. Kim, M. Kilgour, and D. Segal, "Intermediate Coherent-Incoherent Charge Transport: DNA as a Case Study," *Journal of Physical Chemistry C*, vol. 120, no. 42, pp. 23951–23962, 2016, doi: 10.1021/acs.jpcc.6b07602.

[140] M. Buttiker, "Coherent and sequential tunneling in series barriers," *IBM J Res Dev*, vol. 32, no. 1, pp. 63–75, Jan. 1988, doi: 10.1147/rd.321.0063.

[141] S. R. Patil *et al.*, "Quantum Transport in DNA Heterostructures: Implications for Nanoelectronics," *ACS Appl Nano Mater*, vol. 4, no. 10, pp. 10029–10037, Oct. 2021, doi: 10.1021/acsanm.1c01087.

[142] R. Gutiérrez, R. A. Caetano, B. P. Woiczikowski, T. Kubar, M. Elstner, and G. Cuniberti, "Charge transport through biomolecular wires in a solvent: Bridging molecular dynamics and model hamiltonian approaches," *Phys Rev Lett*, vol. 102, no. 20, p. 208102, May 2009, doi: 10.1103/PhysRevLett.102.208102.

[143] Y. A. Mantz, F. L. Gervasio, T. Laino, and M. Parrinello, "Solvent Effects on Charge Spatial Extent in DNA and Implications for Transfer," *Phys Rev Lett*, vol. 99, no. 5, p. 058104, Aug. 2007, doi: 10.1103/PhysRevLett.99.058104.

[144] L. M. Westerhoff and K. M. Merz, "Quantum mechanical description of the interactions between DNA and water," *J Mol Graph Model*, vol. 24, no. 6, pp. 440–455, May 2006, doi: 10.1016/j.jmgm.2005.08.010.

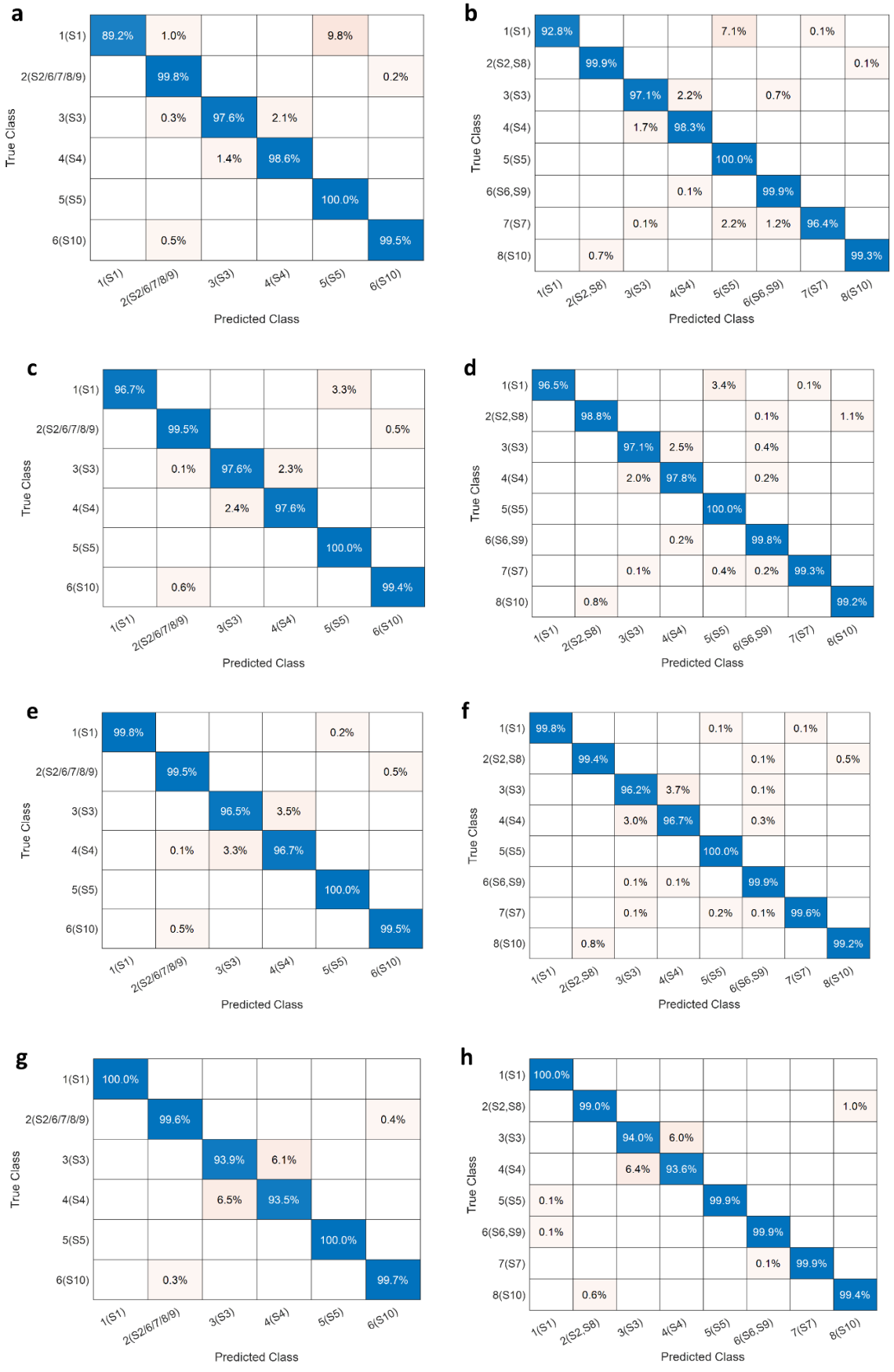
[145] K. Fukuzawa *et al.*, "Explicit solvation modulates intra- and inter-molecular interactions within DNA: Electronic aspects revealed by the ab initio fragment molecular orbital (FMO) method," *Comput Theor Chem*, vol. 1054, pp. 29–37, Feb. 2015, doi: 10.1016/j.comptc.2014.11.020.

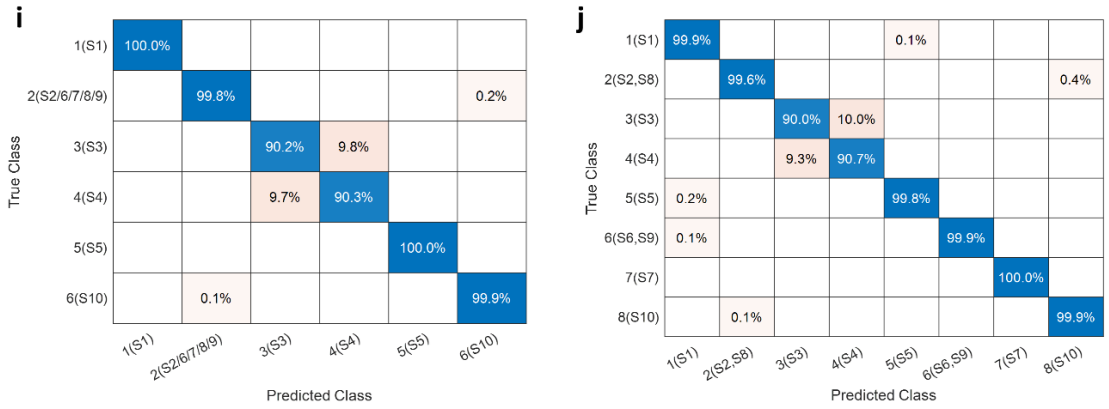
[146] A. Muñoz-Losa, D. Markovitsi, and R. Improta, “A State-Specific PCM–DFT method to include dynamic solvent effects in the calculation of ionization energies: Application to DNA bases,” *Chem Phys Lett*, vol. 634, pp. 20–24, Aug. 2015, doi: 10.1016/J.CPLETT.2015.05.045.

[147] I. Horváth, N. Jeszenői, M. Bálint, G. Paragi, and C. Hetényi, “A Fragmenting Protocol with Explicit Hydration for Calculation of Binding Enthalpies of Target-Ligand Complexes at a Quantum Mechanical Level,” *International Journal of Molecular Sciences 2019, Vol. 20, Page 4384*, vol. 20, no. 18, p. 4384, Sep. 2019, doi: 10.3390/IJMS20184384.

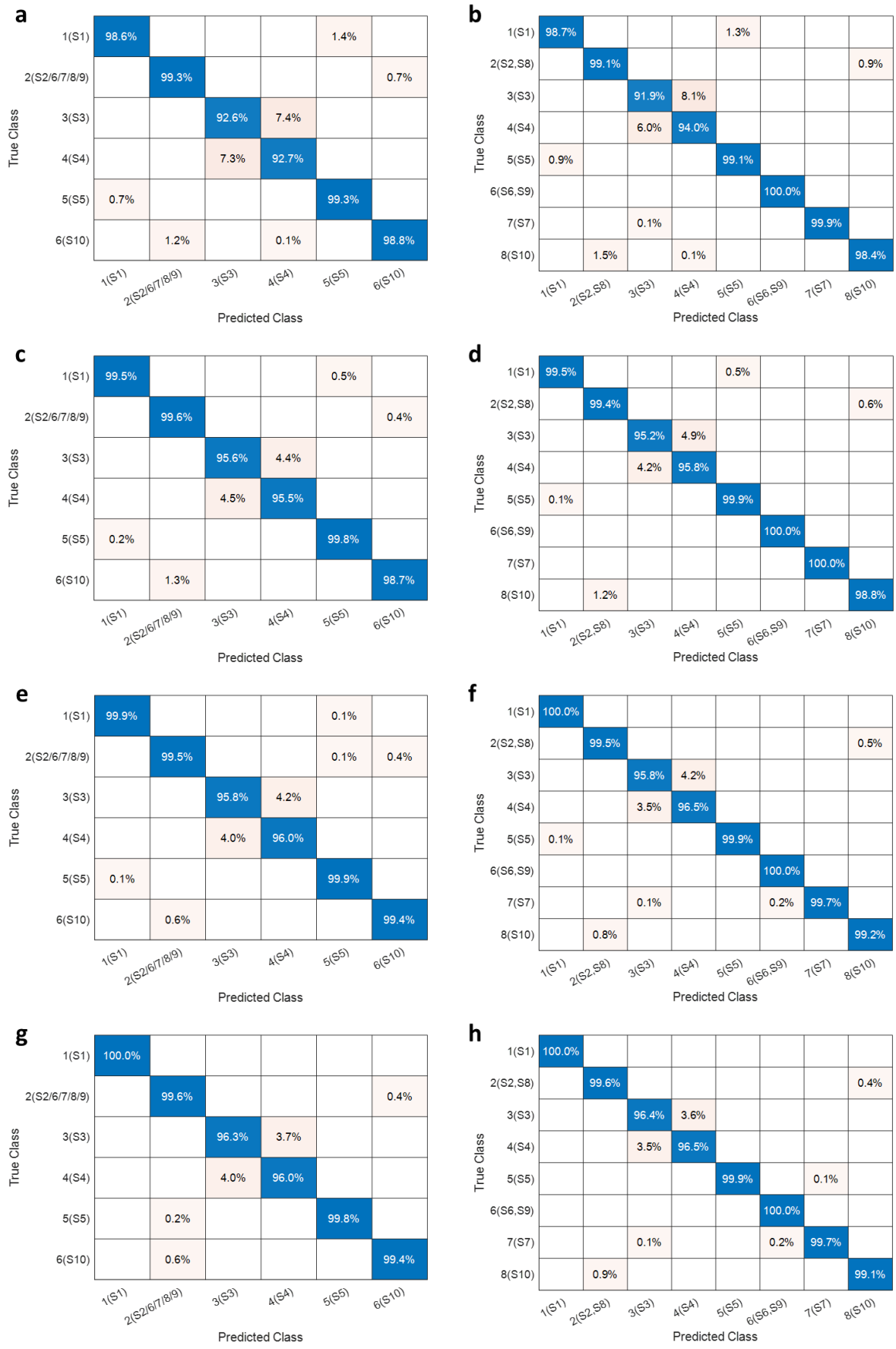
[148] K. Shimamura *et al.*, “Influence of solvating water molecules on the attacking mechanisms of OH-radical to DNA base pairs: DFT calculations in explicit waters,” *Struct Chem*, vol. 27, no. 6, pp. 1793–1806, Dec. 2016, doi: 10.1007/S11224-016-0800-3/FIGURES/13.

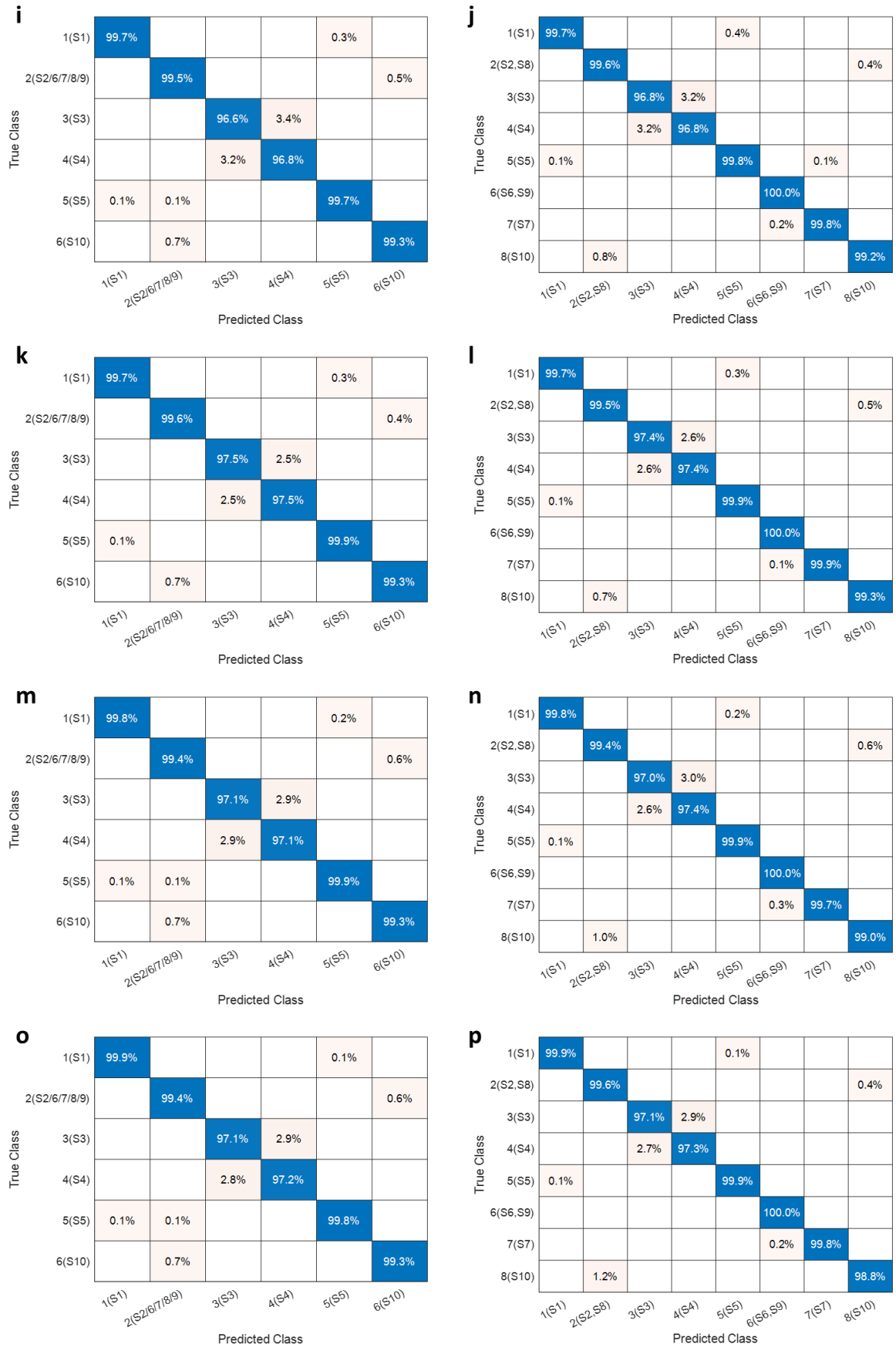
Appendix A Detail Confusion Matrices

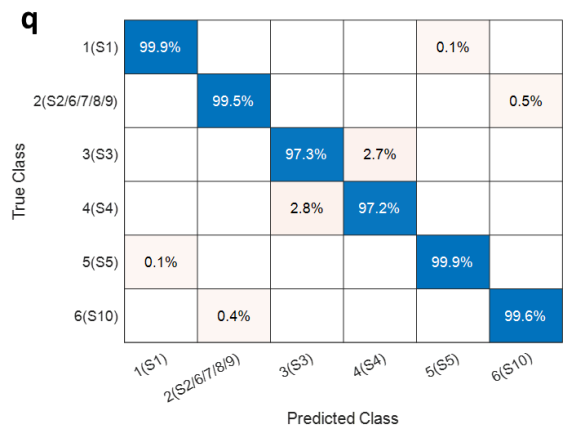


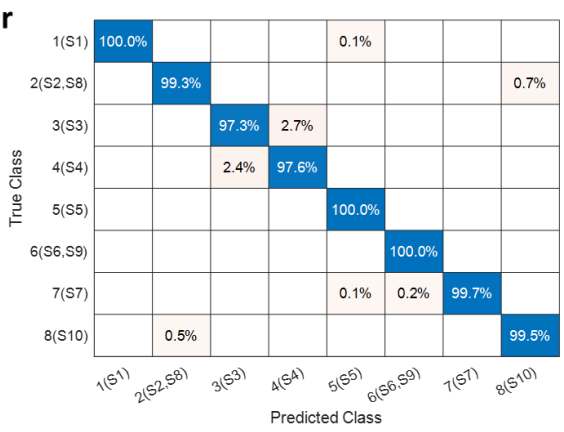


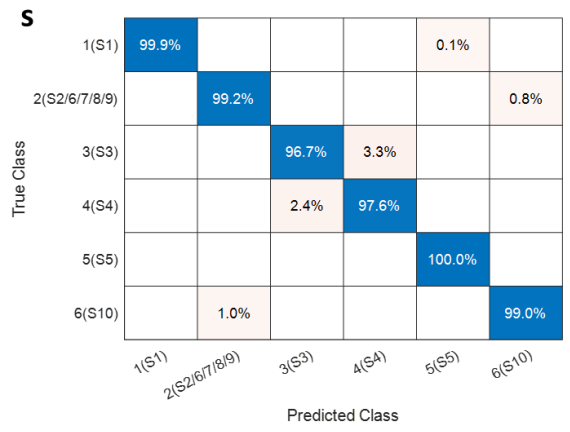
Appendix Figure A-1. Performance analysis of baseline classifiers with respect to the R^2 test threshold parameter, β . The left-column and right-column figures correspond to TLS-1 with 6 classes and TLS-2 with 8 classes respectively. The confusion matrices correspond to baseline parameter values (N_{bins} = 600, H = 30) and changing parameter β. (a), (b) for β = 0.87. (c), (d) for β = 0.90. (e), (f) for β = 0.95 (same as Figure 2-8). (g), (h) for β = 0.99. (i), (j) for β = 1.00

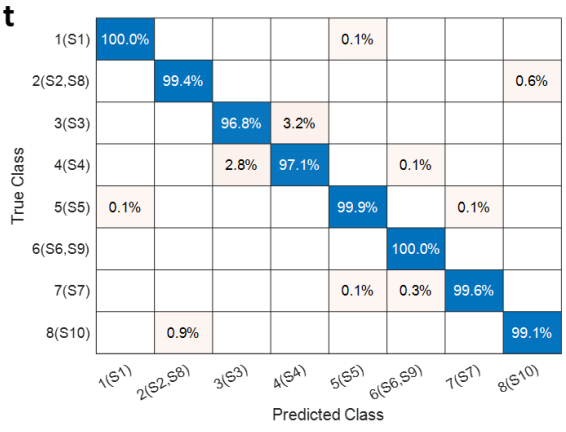

a
1(S1) 98.6% 1.4%
2(S2/6/7/8/9) 99.3% 0.7%
3(S3) 92.6% 7.4%
4(S4) 7.3% 92.7%
5(S5) 0.7% 99.3%
6(S10) 1.2% 0.1% 98.8%
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
True Class
Predicted Class
b
1(S1) 98.7% 1.3%
2(S2,S8) 99.1% 0.9%
3(S3) 91.9% 8.1%
4(S4) 6.0% 94.0%
5(S5) 0.9% 99.1%
6(S6,S9) 100.0%
7(S7) 0.1% 99.9%
8(S10) 1.5% 0.1% 98.4%
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
True Class
Predicted Class
c
1(S1) 99.5% 0.5%
2(S2/6/7/8/9) 99.6% 0.4%
3(S3) 95.6% 4.4%
4(S4) 4.5% 95.5%
5(S5) 0.2% 99.8%
6(S10) 1.3% 98.7%
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
True Class
Predicted Class
d
1(S1) 99.5% 0.5%
2(S2,S8) 99.4% 0.6%
3(S3) 95.2% 4.9%
4(S4) 4.2% 95.8%
5(S5) 0.1% 99.9%
6(S6,S9) 100.0%
7(S7) 100.0%
8(S10) 1.2% 98.8%
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
True Class
Predicted Class
e
1(S1) 99.9% 0.1%
2(S2/6/7/8/9) 99.5% 0.1% 0.4%
3(S3) 95.8% 4.2%
4(S4) 4.0% 96.0%
5(S5) 0.1% 99.9%
6(S10) 0.6% 99.4%
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
True Class
Predicted Class
f
1(S1) 100.0%
2(S2,S8) 99.5% 0.5%
3(S3) 95.8% 4.2%
4(S4) 3.5% 96.5%
5(S5) 0.1% 99.9%
6(S6,S9) 100.0%
7(S7) 0.1% 0.2% 99.7%
8(S10) 0.8% 99.2%
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
True Class
Predicted Class
g
1(S1) 100.0%
2(S2/6/7/8/9) 99.6% 0.4%
3(S3) 96.3% 3.7%
4(S4) 4.0% 96.0%
5(S5) 0.2% 99.8%
6(S10) 0.6% 99.4%
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
True Class
Predicted Class
h
1(S1) 100.0%
2(S2,S8) 99.6% 0.4%
3(S3) 96.4% 3.6%
4(S4) 3.5% 96.5%
5(S5) 99.9% 0.1%
6(S6,S9) 100.0%
7(S7) 0.1% 0.2% 99.7%
8(S10) 0.9% 99.1%
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
True Class
Predicted Class

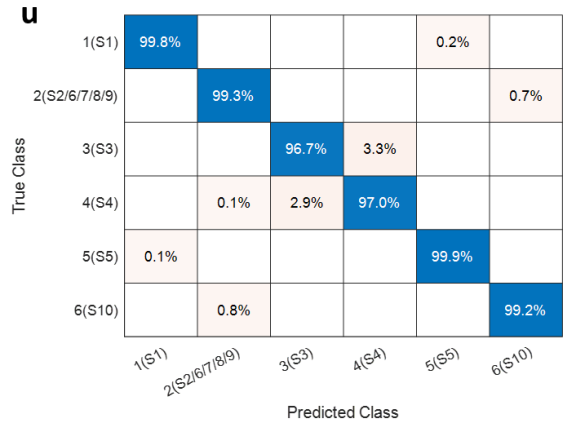

i
1(S1) 99.7% 0.3%
2(S2/6/7/8/9) 99.5% 0.5%
3(S3) 96.6% 3.4%
4(S4) 3.2% 96.8%
5(S5) 0.1% 0.1% 99.7%
6(S10) 0.7% 99.3%
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
True Class
Predicted Class
j
1(S1) 99.7% 0.4%
2(S2,S8) 99.6% 0.4%
3(S3) 96.8% 3.2%
4(S4) 3.2% 96.8%
5(S5) 0.1% 99.8% 0.1%
6(S6,S9) 100.0%
7(S7) 0.2% 99.8%
8(S10) 0.8% 99.2%
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
True Class
Predicted Class
k
1(S1) 99.7% 0.3%
2(S2/6/7/8/9) 99.6% 0.4%
3(S3) 97.5% 2.5%
4(S4) 2.5% 97.5%
5(S5) 0.1% 99.9%
6(S10) 0.7% 99.3%
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
True Class
Predicted Class
l
1(S1) 99.7% 0.3%
2(S2,S8) 99.5% 0.5%
3(S3) 97.4% 2.6%
4(S4) 2.6% 97.4%
5(S5) 0.1% 99.9%
6(S6,S9) 100.0%
7(S7) 0.1% 99.9%
8(S10) 0.7% 99.3%
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
True Class
Predicted Class
m
1(S1) 99.8% 0.2%
2(S2/6/7/8/9) 99.4% 0.6%
3(S3) 97.1% 2.9%
4(S4) 2.9% 97.1%
5(S5) 0.1% 0.1% 99.9%
6(S10) 0.7% 99.3%
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
True Class
Predicted Class
n
1(S1) 99.8% 0.2%
2(S2,S8) 99.4% 0.6%
3(S3) 97.0% 3.0%
4(S4) 2.6% 97.4%
5(S5) 0.1% 99.9%
6(S6,S9) 100.0%
7(S7) 0.3% 99.7%
8(S10) 1.0% 99.0%
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
True Class
Predicted Class
o
1(S1) 99.9% 0.1%
2(S2/6/7/8/9) 99.4% 0.6%
3(S3) 97.1% 2.9%
4(S4) 2.8% 97.2%
5(S5) 0.1% 0.1% 99.8%
6(S10) 0.7% 99.3%
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
True Class
Predicted Class
p
1(S1) 99.9% 0.1%
2(S2,S8) 99.6% 0.4%
3(S3) 97.1% 2.9%
4(S4) 2.7% 97.3%
5(S5) 0.1% 99.9%
6(S6,S9) 100.0%
7(S7) 0.2% 99.8%
8(S10) 1.2% 98.8%
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
True Class
Predicted Class

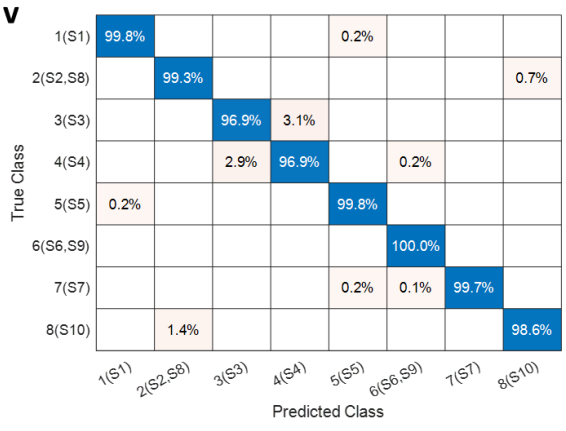

q
1(S1) 99.9% 0.1%
2(S2/6/7/8/9) 99.5% 0.5%
3(S3) 97.3% 2.7%
4(S4) 2.8% 97.2%
5(S5) 0.1% 99.9%
6(S10) 0.4% 99.6%
True Class
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
Predicted Class

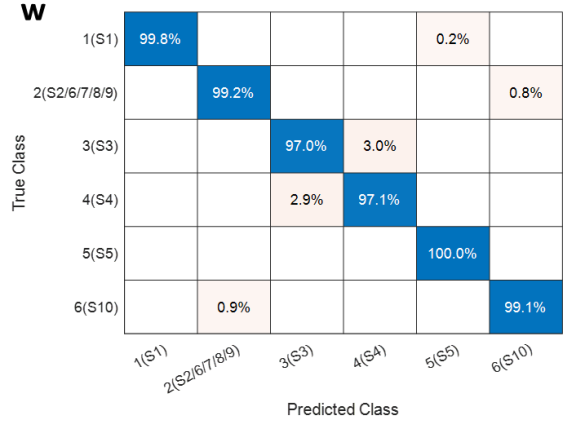

r
1(S1) 100.0% 0.1%
2(S2,S8) 99.3% 0.7%
3(S3) 97.3% 2.7%
4(S4) 2.4% 97.6%
5(S5) 100.0%
6(S6,S9) 100.0%
7(S7) 0.1% 0.2% 99.7%
8(S10) 0.5% 99.5%
True Class
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
Predicted Class

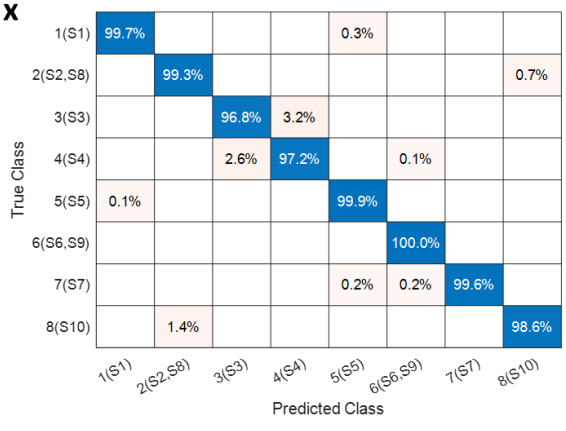

s
1(S1) 99.9% 0.1%
2(S2/6/7/8/9) 99.2% 0.8%
3(S3) 96.7% 3.3%
4(S4) 2.4% 97.6%
5(S5) 100.0%
6(S10) 1.0% 99.0%
True Class
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
Predicted Class

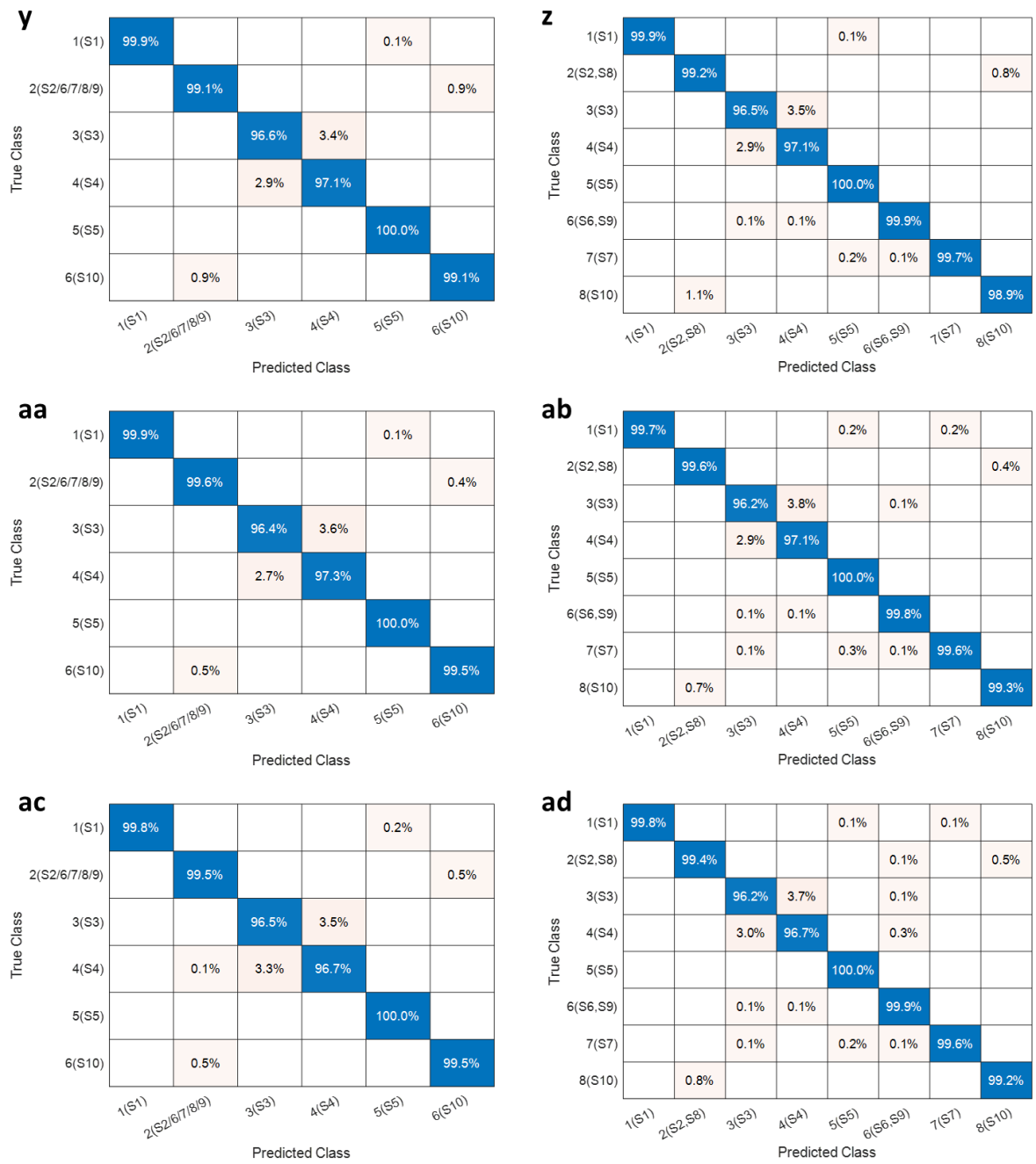

t
1(S1) 100.0% 0.1%
2(S2,S8) 99.4% 0.6%
3(S3) 96.8% 3.2%
4(S4) 2.8% 97.1% 0.1%
5(S5) 0.1% 99.9% 0.1%
6(S6,S9) 100.0%
7(S7) 0.1% 0.3% 99.6%
8(S10) 0.9% 99.1%
True Class
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
Predicted Class


u
1(S1) 99.8% 0.2%
2(S2/6/7/8/9) 99.3% 0.7%
3(S3) 96.7% 3.3%
4(S4) 0.1% 2.9% 97.0%
5(S5) 0.1% 99.9%
6(S10) 0.8% 99.2%
True Class
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
Predicted Class

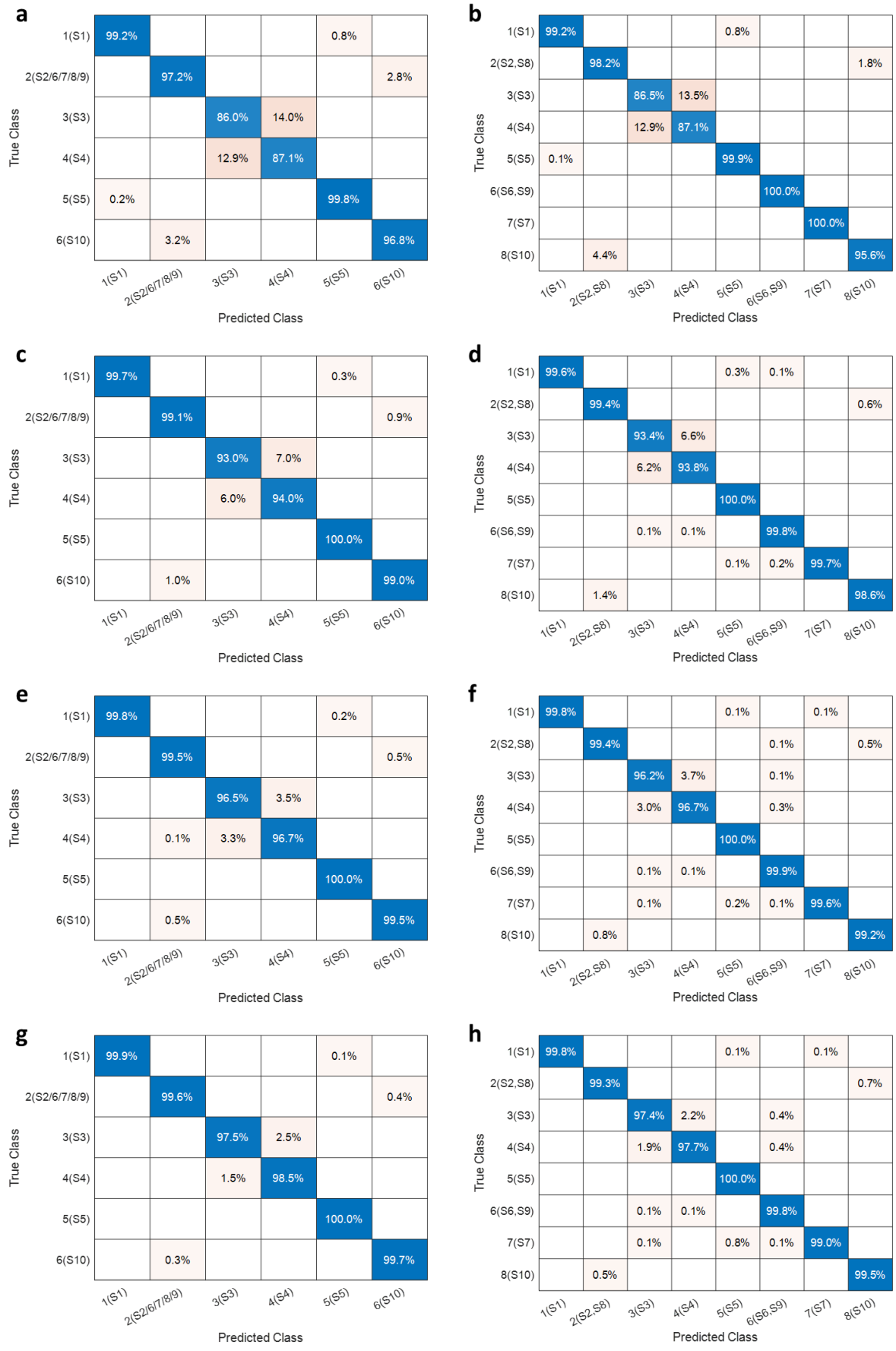

v
1(S1) 99.8% 0.2%
2(S2,S8) 99.3% 0.7%
3(S3) 96.9% 3.1%
4(S4) 2.9% 96.9% 0.2%
5(S5) 0.2% 99.8%
6(S6,S9) 100.0%
7(S7) 0.2% 0.1% 99.7%
8(S10) 1.4% 98.6%
True Class
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
Predicted Class

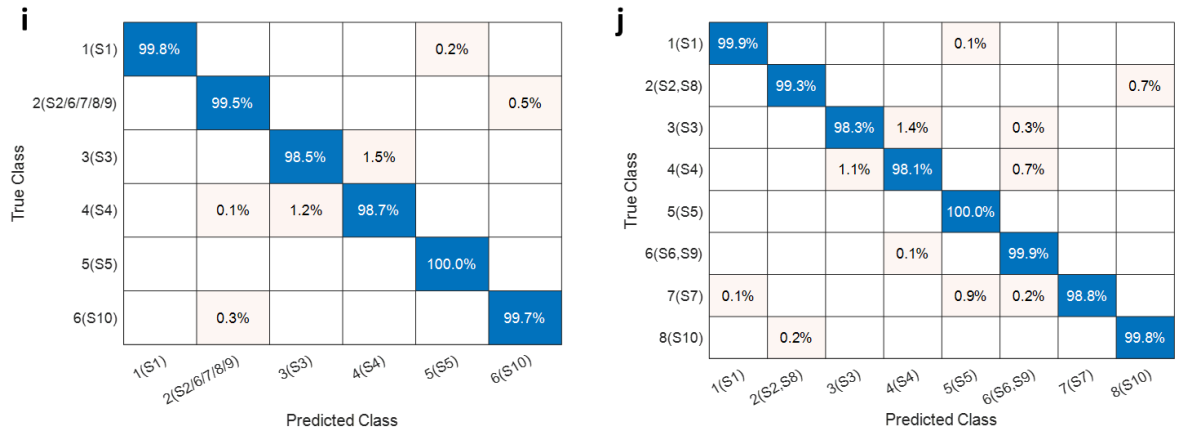

w
1(S1) 99.8% 0.2%
2(S2/6/7/8/9) 99.2% 0.8%
3(S3) 97.0% 3.0%
4(S4) 2.9% 97.1%
5(S5) 100.0%
6(S10) 0.9% 99.1%
True Class
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
Predicted Class


x
1(S1) 99.7% 0.3%
2(S2,S8) 99.3% 0.7%
3(S3) 96.8% 3.2%
4(S4) 2.6% 97.2% 0.1%
5(S5) 0.1% 99.9%
6(S6,S9) 100.0%
7(S7) 0.2% 0.2% 99.6%
8(S10) 1.4% 98.6%
True Class
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
Predicted Class



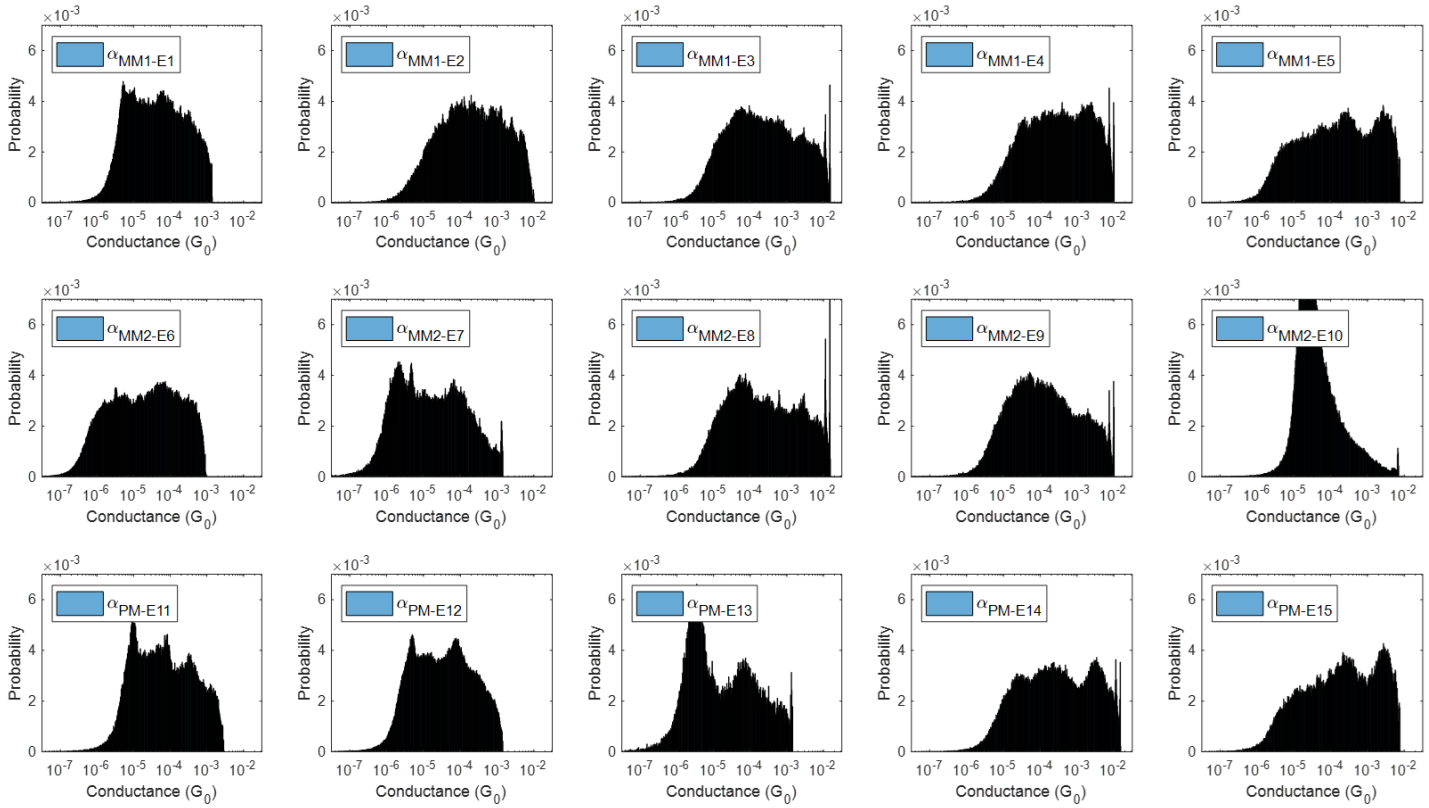
Appendix Figure A-2. Performance analysis of baseline classifiers with respect to number of histogram bins, N_{bins}. The left-column and right-column figures correspond to TLS-1 with 6 classes and TLS-2 with 8 classes respectively. The confusion matrices correspond to baseline parameter values (β = 0.95, H = 30) and changing parameter N_{bins}. (a), (b) for N_{bins} = 10. (c), (d) for N_{bins} = 20. (e), (f) for N_{bins} = 30. (g), (h) for N_{bins} = 40. (i), (j) for N_{bins} = 50. (k), (l) for N_{bins} = 60. (m), (n) for N_{bins} = 75. (o), (p) for N_{bins} = 100. (q), (r) for N_{bins} = 120. (s), (t) for N_{bins} = 150. (u), (v) for N_{bins} = 200. (w), (x) for N_{bins} = 300. (y), (z) for N_{bins} = 400. (aa), (ab) for N_{bins} = 500. (ac), (ad) for N_{bins} = 600 (same as Figure 2-8).


a
1(S1) 99.2% 0.8%
2(S2/6/7/8/9) 97.2% 2.8%
3(S3) 86.0% 14.0%
4(S4) 12.9% 87.1%
5(S5) 0.2% 99.8%
6(S10) 3.2% 96.8%
True Class
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
Predicted Class
b
1(S1) 99.2% 0.8%
2(S2,S8) 98.2% 1.8%
3(S3) 86.5% 13.5%
4(S4) 12.9% 87.1%
5(S5) 0.1% 99.9%
6(S6,S9) 100.0%
7(S7) 100.0%
8(S10) 4.4% 95.6%
True Class
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
Predicted Class
c
1(S1) 99.7% 0.3%
2(S2/6/7/8/9) 99.1% 0.9%
3(S3) 93.0% 7.0%
4(S4) 6.0% 94.0%
5(S5) 100.0%
6(S10) 1.0% 99.0%
True Class
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
Predicted Class
d
1(S1) 99.6% 0.3% 0.1%
2(S2,S8) 99.4% 0.6%
3(S3) 93.4% 6.6%
4(S4) 6.2% 93.8%
5(S5) 100.0%
6(S6,S9) 0.1% 0.1% 99.8%
7(S7) 0.1% 0.2% 99.7%
8(S10) 1.4% 98.6%
True Class
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
Predicted Class
e
1(S1) 99.8% 0.2%
2(S2/6/7/8/9) 99.5% 0.5%
3(S3) 96.5% 3.5%
4(S4) 0.1% 3.3% 96.7%
5(S5) 100.0%
6(S10) 0.5% 99.5%
True Class
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
Predicted Class
f
1(S1) 99.8% 0.1% 0.1%
2(S2,S8) 99.4% 0.1% 0.5%
3(S3) 96.2% 3.7% 0.1%
4(S4) 3.0% 96.7% 0.3%
5(S5) 100.0%
6(S6,S9) 0.1% 0.1% 99.9%
7(S7) 0.1% 0.2% 0.1% 99.6%
8(S10) 0.8% 99.2%
True Class
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
Predicted Class
g
1(S1) 99.9% 0.1%
2(S2/6/7/8/9) 99.6% 0.4%
3(S3) 97.5% 2.5%
4(S4) 1.5% 98.5%
5(S5) 100.0%
6(S10) 0.3% 99.7%
True Class
1(S1) 2(S2/6/7/8/9) 3(S3) 4(S4) 5(S5) 6(S10)
Predicted Class
h
1(S1) 99.8% 0.1% 0.1%
2(S2,S8) 99.3% 0.7%
3(S3) 97.4% 2.2% 0.4%
4(S4) 1.9% 97.7% 0.4%
5(S5) 100.0%
6(S6,S9) 0.1% 0.1% 99.8%
7(S7) 0.1% 0.8% 0.1% 99.0%
8(S10) 0.5% 99.5%
True Class
1(S1) 2(S2,S8) 3(S3) 4(S4) 5(S5) 6(S6,S9) 7(S7) 8(S10)
Predicted Class

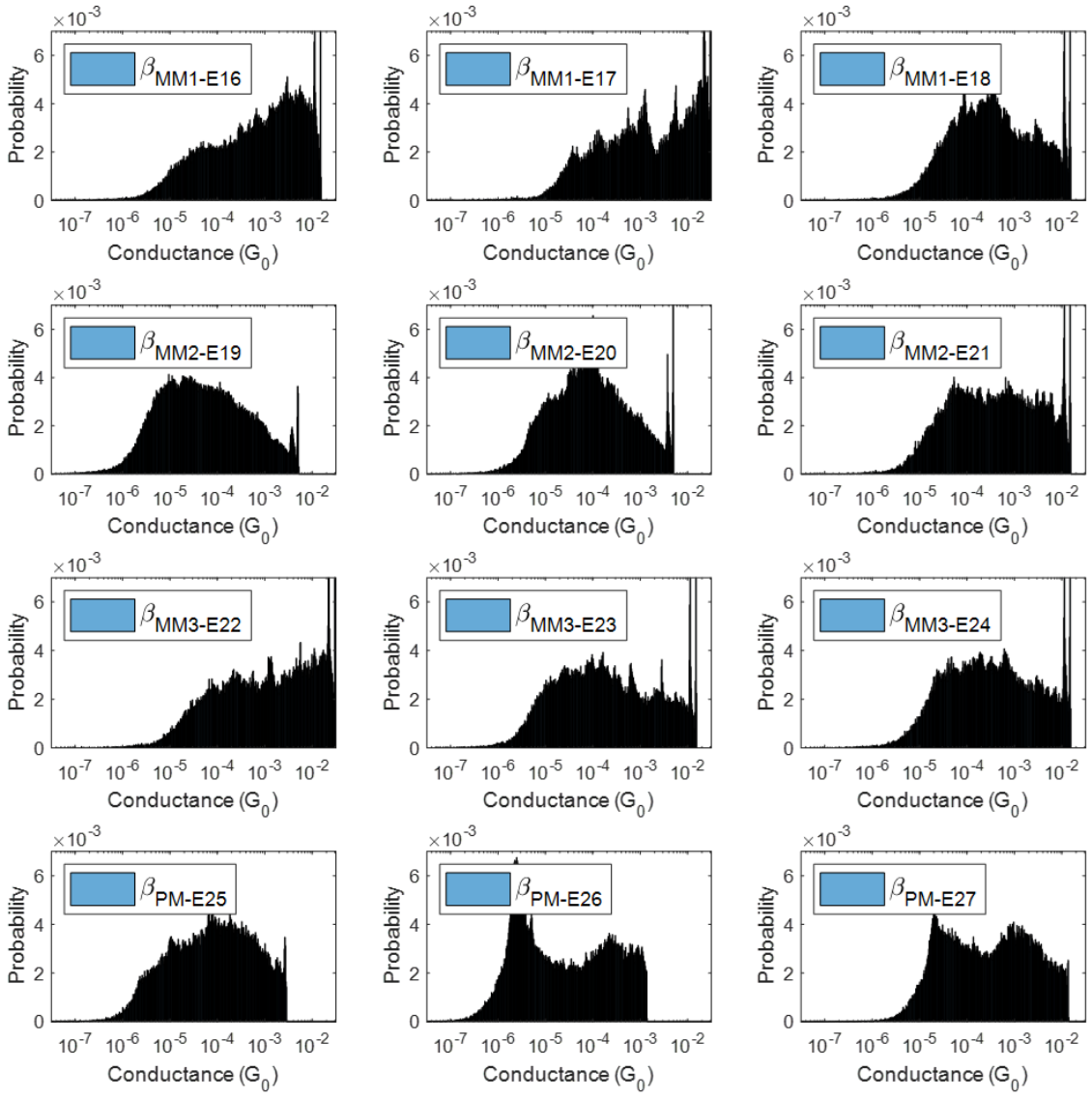


Appendix Figure A-3. Performance analysis of baseline classifiers with respect to the number of traces used to compute a conductance histogram, *H*. The left-column and right-column figures correspond to TLS-1 with 6 classes and TLS-2 with 8 classes respectively. The confusion matrices correspond to baseline parameter values (β = 0.95, N_{bins} = 600) and changing parameter *H*. (a), (b) for *H* = 10. (c), (d) for *H* = 20. (e), (f) for *H* = 30 (same as Figure 2-8). (g), (h) for *H* = 40. (i), (j) for *H* = 50.

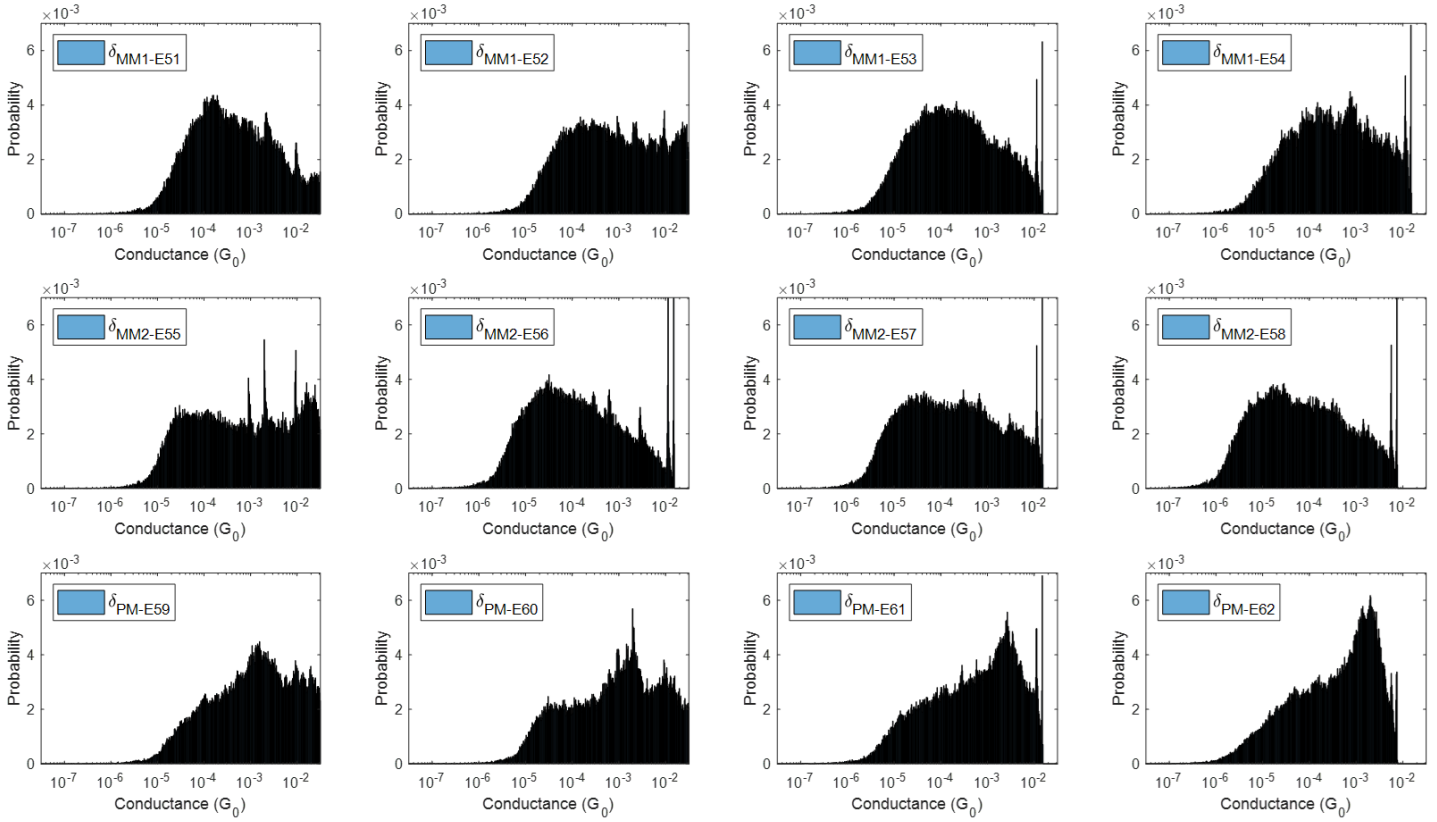
Appendix B 1D and 2D Conductance Histogram of COVID-19



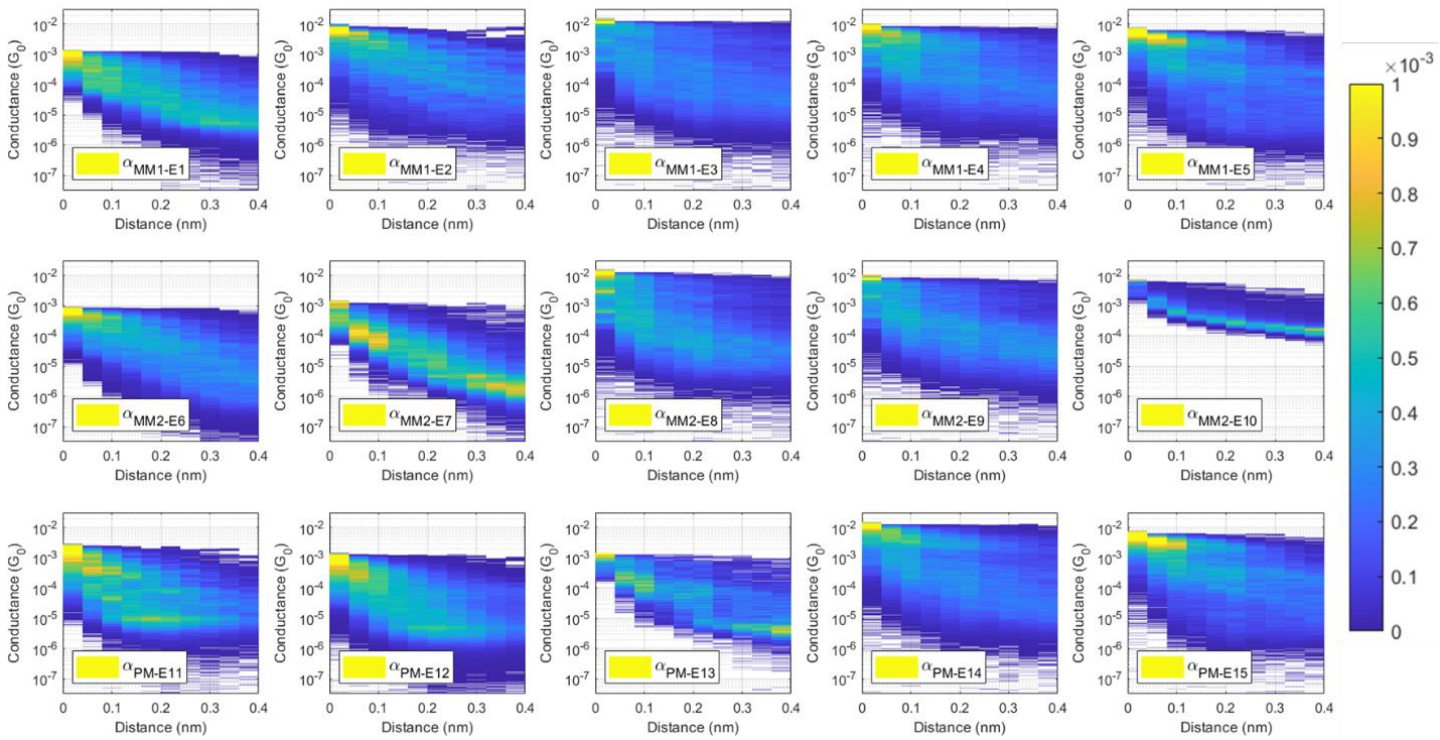
Appendix Figure B-1. Empirical large sample 1D conductance histograms for Alpha variant, with exponential fitting, R^2 test ($\beta = 0.95$), and low pass filter.



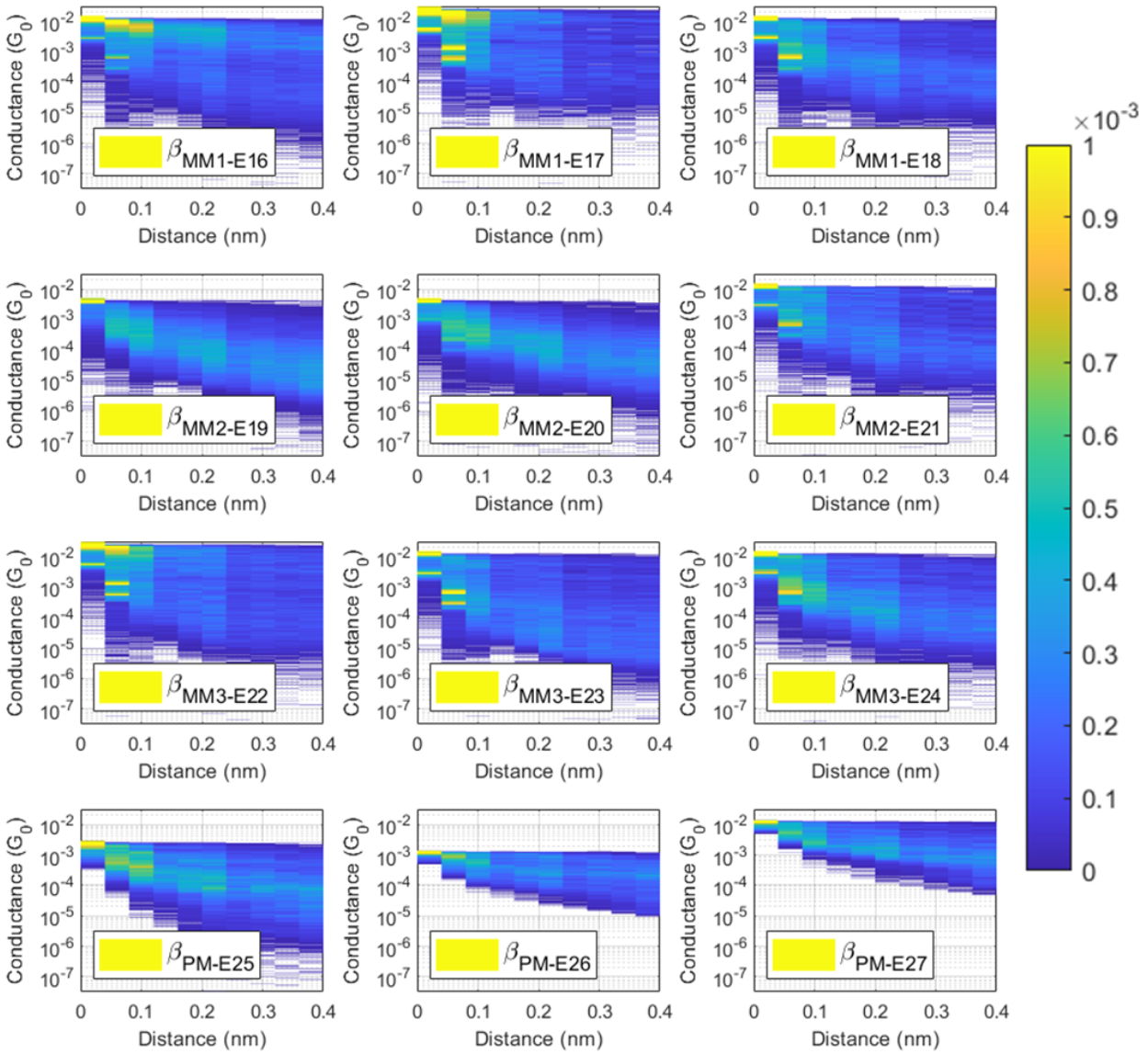
Appendix Figure B-2. Empirical large sample 1D conductance histograms for Beta variant, with exponential fitting, R^2 test ($\beta = 0.95$), and low pass filter.



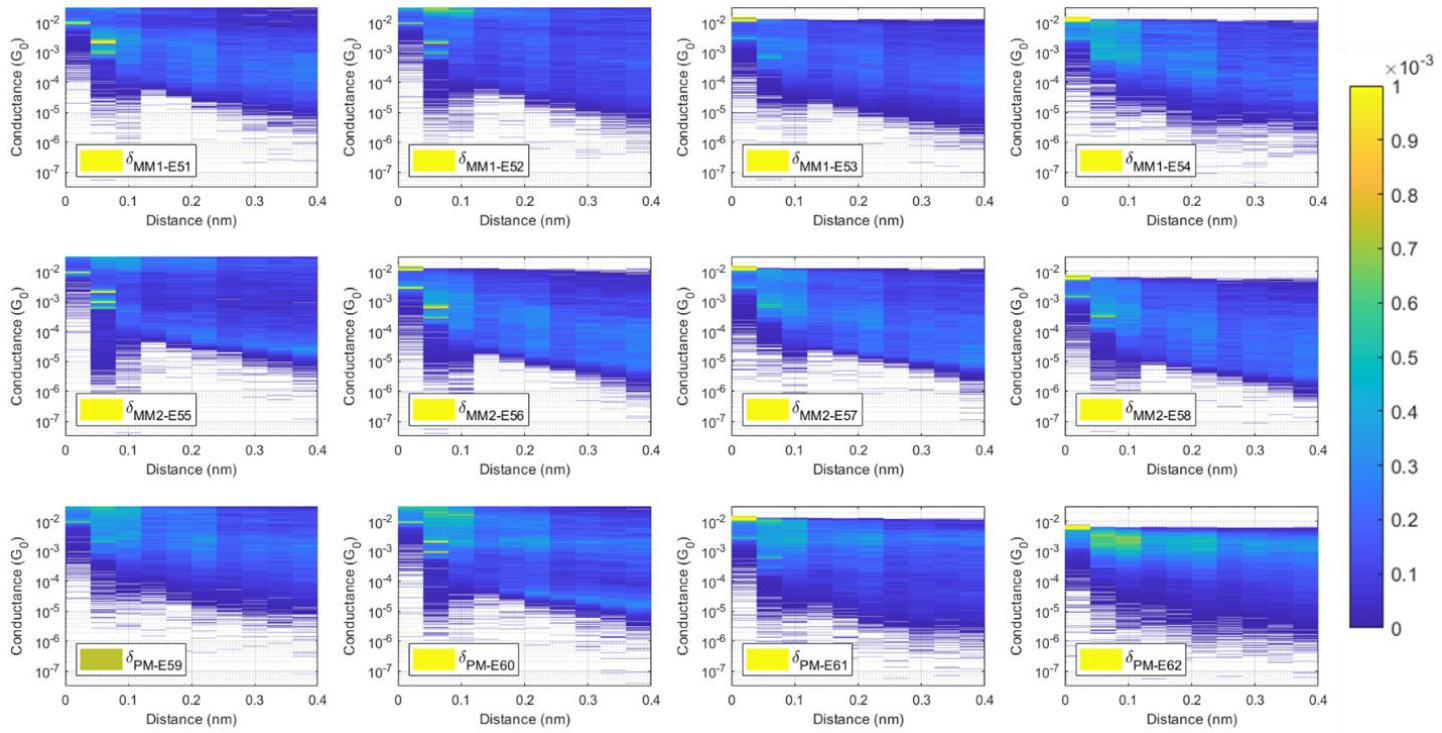
Appendix Figure B-3. Empirical large sample 1D conductance histograms for Delta variant, with exponential fitting, R^2 test ($\beta = 0.95$), and low pass filter.



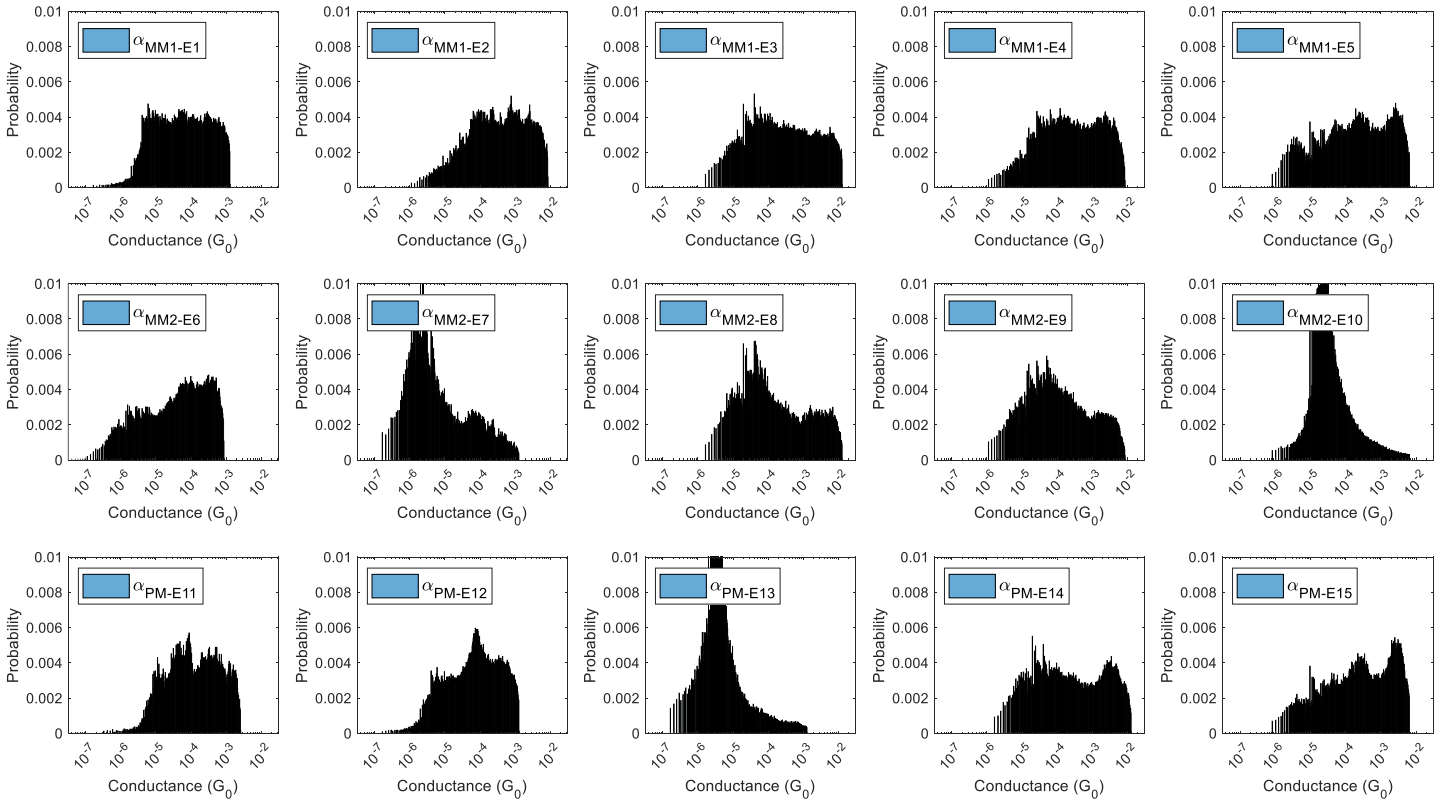
Appendix Figure B-4. Empirical large sample 2D conductance histograms for Alpha variant, with exponential fitting, R^2 test ($\beta = 0.95$), and low pass filter. For better visualization, all elements which have a value larger than 0.001 are represented using the same color scheme as the value 0.001.



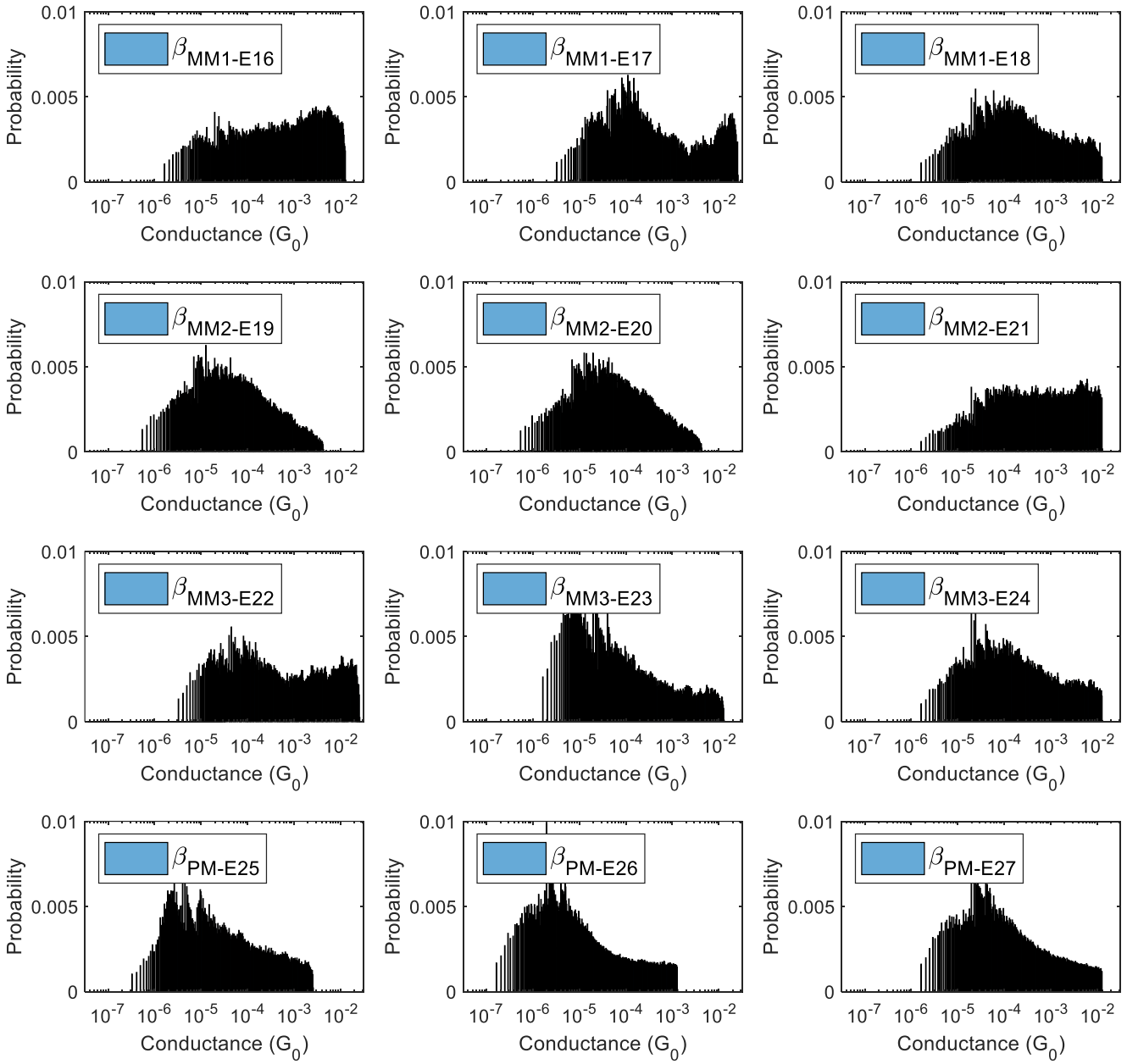
Appendix Figure B-5. Empirical large sample 2D conductance histograms for Beta variant, with exponential fitting, R^2 test ($\beta = 0.95$), and low pass filter. For better visualization, all elements which have a value larger than 0.001 are represented using the same color scheme as the value 0.001.



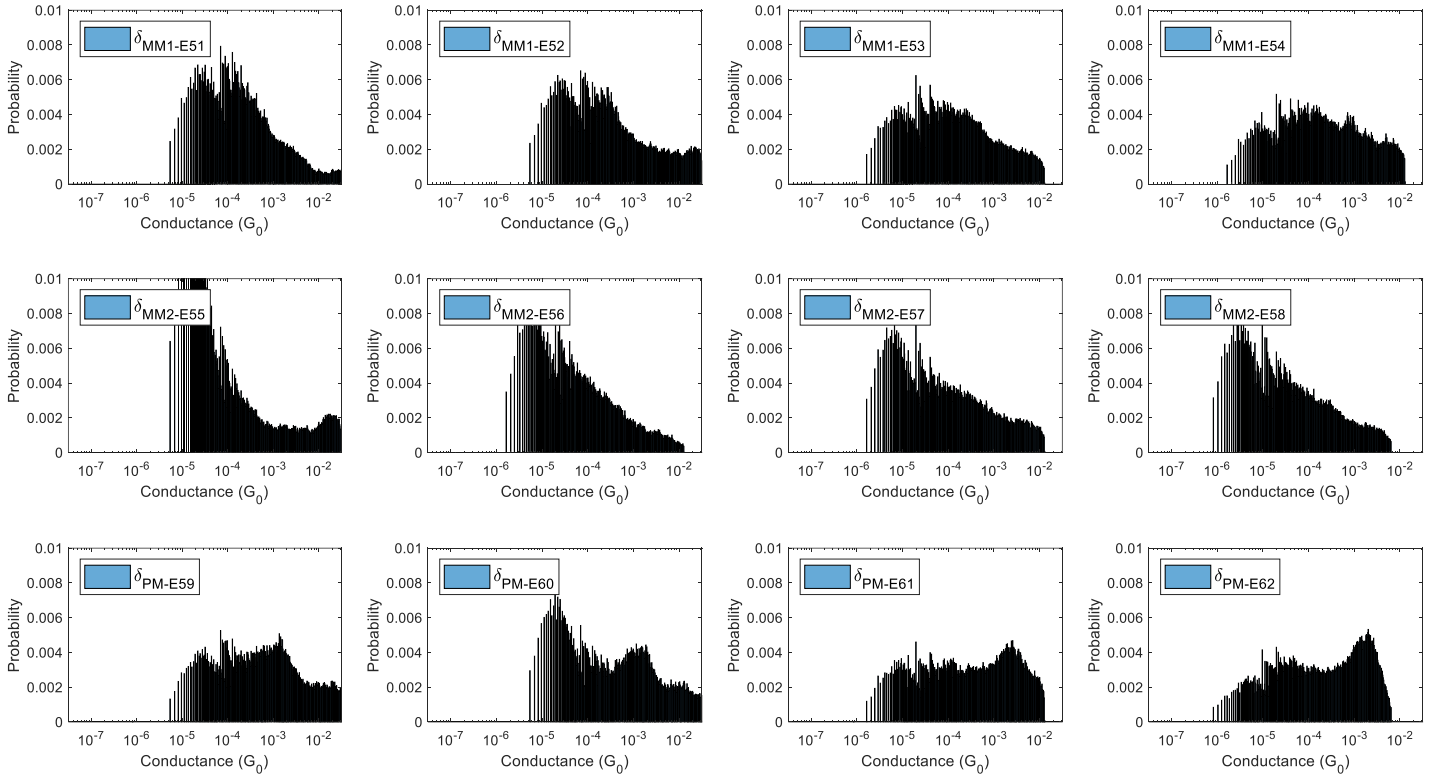
Appendix Figure B-6. Empirical large sample 2D conductance histograms for Delta variant, with exponential fitting, R^2 test ($\beta = 0.95$), and low pass filter. For better visualization, all elements which have a value larger than 0.001 are represented using the same color scheme as the value 0.001.



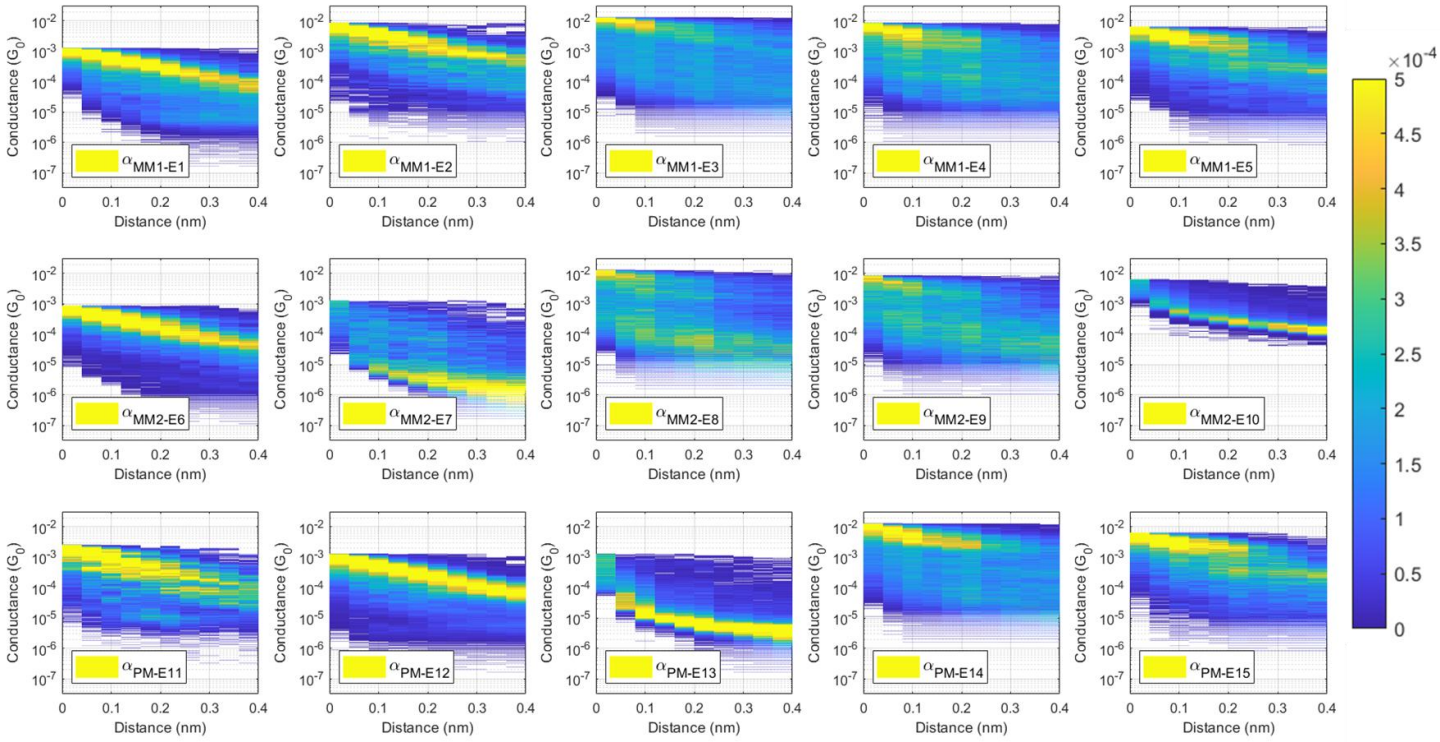
Appendix Figure B-7. Empirical large sample 1D conductance histograms for Alpha variant, Piecewise Linear Approximation method, with $\beta = 0.7$ and $A_{co} = 4$.



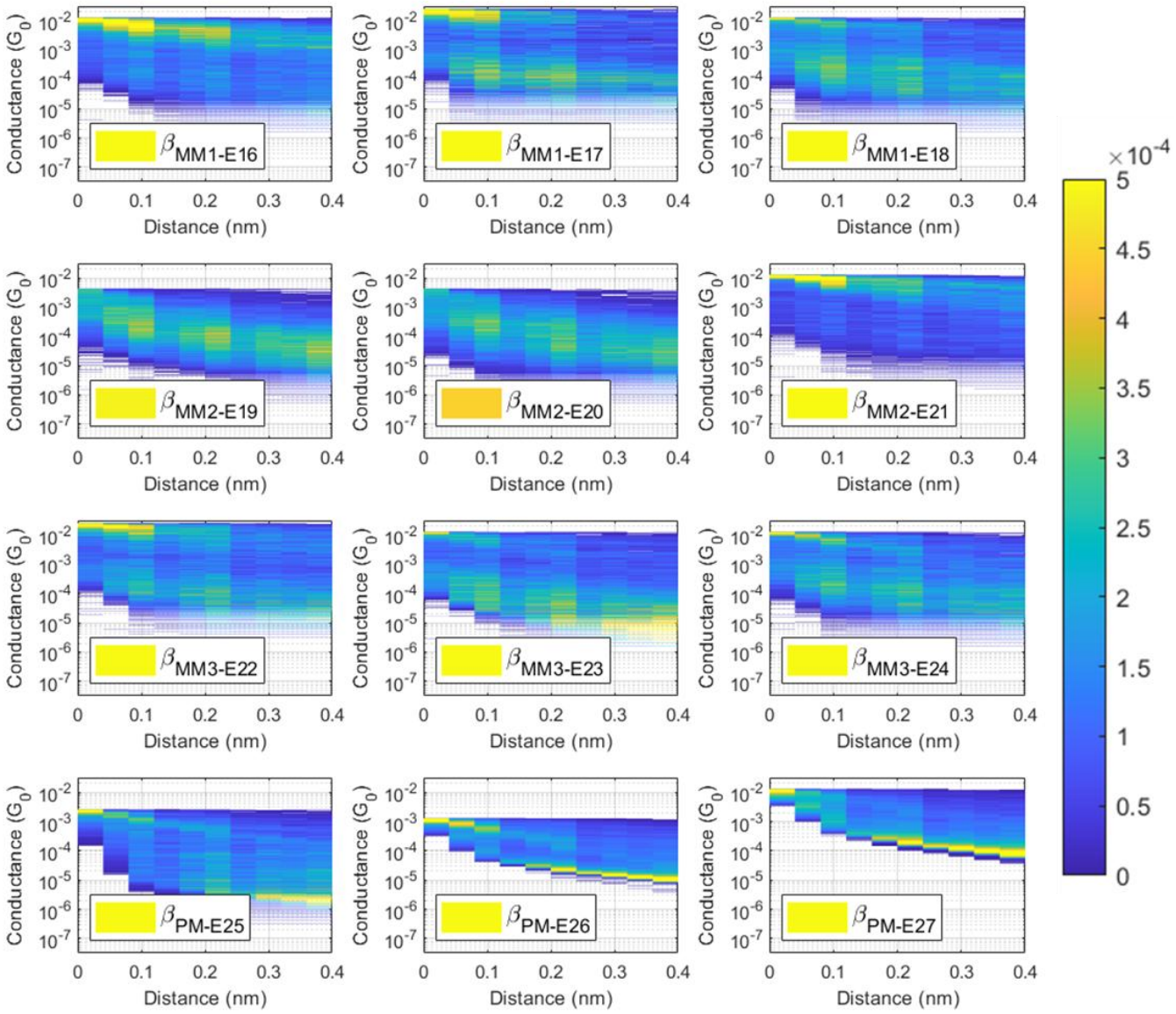
Appendix Figure B-8. Empirical large sample 1D conductance histograms for Beta variant, Piecewise Linear Approximation method, with $\beta = 0.7$ and $A_{co} = 4$.



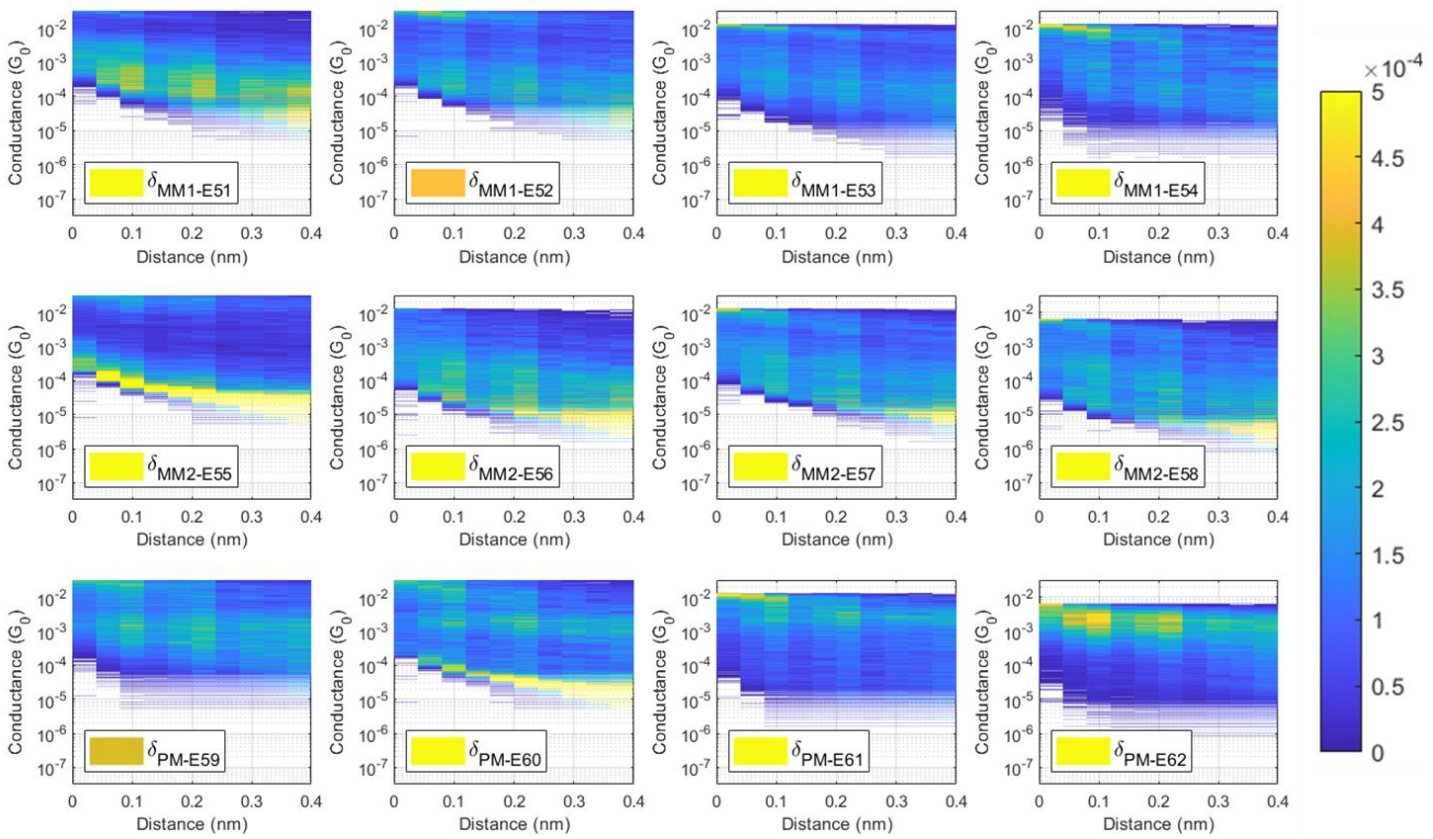
Appendix Figure B-9. Empirical large sample 1D conductance histograms for Delta variant, Piecewise Linear Approximation method, with $\beta = 0.7$ and $A_{co} = 4$.



Appendix Figure B-10. Empirical large sample 2D conductance histograms for Alpha variant, Piecewise Linear Approximation method, with $\beta = 0.7$ and $A_{co} = 4$. For better visualization, all elements which have a value larger than 0.0005 are represented using the same color scheme as the value 0.0005.



Appendix Figure B-11. Empirical large sample 2D conductance histograms for Beta variant, Piecewise Linear Approximation method, with $\beta = 0.7$ and $A_{co} = 4$. For better visualization, all elements which have a value larger than 0.0005 are represented using the same color scheme as the value 0.0005.



Appendix Figure B-12. Empirical large sample 2D conductance histograms for Delta variant, Piecewise Linear Approximation method, with $\beta = 0.7$ and $A_{co} = 4$. For better visualization, all elements which have a value larger than 0.0005 are represented using the same color scheme as the value 0.0005.

Appendix C Coordinates of the DNA Molecule

We list each atom's (x, y, z) coordinates for our modeling system below. Sodium ions are located at the end of the list. The sodium ion highlighted in bold is the 7th Na^+ ion we removed in Section 5 (see Figure 5-5).

H	-0.4230	-8.1500	-2.0940
O	0.4270	-7.8260	-1.7880
C	1.4430	-7.5100	-2.7560
H	1.1000	-6.7450	-3.4530
H	1.6740	-8.4230	-3.3050
C	2.6950	-7.0200	-2.0530
H	3.5740	-7.0560	-2.6960
O	2.4770	-5.6300	-1.8230
C	2.3080	-5.3330	-0.4520
H	3.1120	-4.6850	-0.1020
N	1.0980	-4.4680	-0.3700
C	-0.1510	-5.0050	-0.3990
H	-0.2760	-6.0740	-0.4820
C	-1.2540	-4.2170	-0.3250
H	-2.2470	-4.6650	-0.3500
C	-1.0350	-2.8050	-0.2150
N	-2.0660	-1.9770	-0.1380
H	-1.8700	-0.9890	-0.0620
H	-3.0100	-2.3360	-0.1560
N	0.1990	-2.2870	-0.1870
C	1.2890	-3.0930	-0.2630
O	2.4450	-2.6510	-0.2410
C	2.9500	-7.6070	-0.6650
H	2.5860	-8.6290	-0.5540
C	2.1480	-6.6790	0.2470
H	1.3440	-7.2410	0.7220
H	2.8040	-6.2670	1.0140
O	4.3240	-7.5450	-0.3070
P	4.7280	-7.6260	1.2390
O	6.0090	-8.3560	1.3800
O	3.5920	-8.1640	2.0210
O	4.9450	-6.0810	1.5920
C	5.5820	-5.2270	0.6240
H	4.8550	-4.8090	-0.0720
H	6.3060	-5.8290	0.0750
C	6.3060	-4.0950	1.3270
H	7.0380	-3.6070	0.6840
O	5.3130	-3.0990	1.5570
C	5.0020	-2.9580	2.9280
H	5.2720	-1.9610	3.2780
N	3.5150	-2.9690	3.0100
C	2.8190	-4.1380	2.9810
H	3.3460	-5.0770	2.8980
C	1.4640	-4.1480	3.0550
H	0.9240	-5.0950	3.0300
C	0.8110	-2.8780	3.1650
N	-0.5090	-2.8130	3.2420
H	-0.9300	-1.8980	3.3190
H	-1.0620	-3.6580	3.2240
N	1.5060	-1.7330	3.1930
C	2.8610	-1.7440	3.1170
O	3.5360	-0.7080	3.1390
C	6.8580	-4.4200	2.7150
H	7.1640	-5.4600	2.8260
C	5.6640	-4.1410	3.6270
H	5.3440	-5.0690	4.1020
H	5.9520	-3.4220	4.3940
O	7.9330	-3.5620	3.0730
P	8.3080	-3.3900	4.6190
O	9.7730	-3.2280	4.7600
O	7.7050	-4.4930	5.4010
O	7.5750	-2.0130	4.9720
C	7.5880	-0.9480	4.0040
H	6.7540	-1.0370	3.3080
H	8.5280	-1.0100	3.4550
C	7.5090	0.3940	4.7070
H	7.8150	1.2190	4.0640
O	6.1200	0.6160	4.9370
C	5.7850	0.5470	6.3080
H	5.4170	1.5120	6.6580
N	4.5890	-0.3360	6.3900
C	4.7130	-1.6900	6.3610
H	5.6910	-2.1400	6.2780
C	3.6230	-2.4950	6.4350
H	3.7430	-3.5780	6.4100
C	2.3480	-1.8520	6.5450
N	1.2420	-2.5750	6.6220
H	0.3640	-2.0830	6.6980
H	1.2910	-3.5840	6.6040
N	2.2370	-0.5170	6.5730
C	3.3400	0.2700	6.4970
O	3.2770	1.5060	6.5190
C	8.1460	0.4550	6.0950
H	9.0050	-0.2070	6.2050
C	7.0160	-0.0210	7.0070
H	7.3020	-0.9590	7.4820

H	6.8260	0.7300	7.7740
O	8.5120	1.7810	6.4530
P	8.7140	2.1400	7.9990
O	9.8040	3.1330	8.1400
O	8.8740	0.8940	8.7810
O	7.3120	2.8240	8.3520
C	6.6960	3.6930	7.3840
H	6.0740	3.1300	6.6890
H	7.4920	4.1950	6.8340
C	5.8430	4.7320	8.0870
H	5.6060	5.5790	7.4430
O	4.5890	4.0960	8.3170
C	4.3590	3.8430	9.6880
H	3.4950	4.4080	10.0390
N	3.9100	2.4250	9.7700
C	4.6630	1.2710	9.7560
H	5.7380	1.3280	9.6750
N	3.9540	0.1710	9.8450
C	2.6400	0.6290	9.9240
C	1.4260	-0.0960	10.0370
O	1.2570	-1.3110	10.0930
N	0.3190	0.7640	10.0870
H	-0.5860	0.3460	10.1680
C	0.3790	2.1430	10.0340
N	-0.7910	2.7830	10.0970
H	-1.7090	2.3670	10.1670
H	-0.7570	3.7910	10.0500
N	1.5190	2.8220	9.9280
C	2.6030	2.0030	9.8790
C	6.3230	5.1570	9.4750
H	7.4070	5.1270	9.5860
C	5.6880	4.1070	10.3870
H	6.4710	3.5160	10.8620
H	5.0930	4.6030	11.1540
O	5.8390	6.4440	9.8330
P	5.7920	6.8540	11.3790
O	6.0900	8.2970	11.5200
O	6.6540	5.9390	12.1610
O	4.2550	6.5820	11.7320
C	3.2460	6.9240	10.7640
H	3.0730	6.1030	10.0680
H	3.5950	7.7980	10.2150
C	1.9460	7.2630	11.4670
H	1.2560	7.8090	10.8230
O	1.3050	6.0110	11.6970
C	1.2680	5.6710	13.0680
H	0.2370	5.6200	13.4180
N	1.7380	4.2600	13.1500
C	3.0640	3.9600	13.1210
H	3.7940	4.7520	13.0380
C	3.4930	2.6740	13.1950
H	4.5600	2.4520	13.1700
C	2.4860	1.6610	13.3050
N	2.8330	0.3850	13.3820
H	2.0940	-0.2990	13.4590
H	3.8080	0.1200	13.3640
N	1.1830	1.9680	13.3330
C	0.7750	3.2600	13.2570
O	-0.4190	3.5820	13.2790
C	2.0840	7.8880	12.8550
H	2.9780	8.5010	12.9660
C	2.1880	6.6660	13.7670
H	3.1690	6.6490	14.2420
H	1.4150	6.7170	14.5330
O	0.9360	8.6450	13.2130
P	0.6570	8.9490	14.7590
O	0.0500	10.2920	14.9000
O	1.8920	8.7160	15.5410
O	-0.4270	7.8260	15.1120
C	-1.4430	7.5100	14.1440
H	-1.1000	6.7450	13.4470
H	-1.6740	8.4230	13.5950
C	-2.6950	7.0200	14.8470
H	-3.5740	7.0560	14.2040
O	-2.4770	5.6300	15.0770
C	-2.3080	5.3330	16.4480
H	-3.1120	4.6860	16.7990
N	-1.0980	4.4680	16.5300
C	0.2320	4.8270	16.5160
H	0.5090	5.8680	16.4350
N	1.0590	3.8140	16.6050
C	0.2180	2.7050	16.6840
C	0.5320	1.3260	16.7970
O	1.6350	0.7900	16.8530
N	-0.6280	0.5400	16.8470
H	-0.5090	-0.4500	16.9290
C	-1.9210	1.0230	16.7940
N	-2.8910	0.1080	16.8570
H	-2.7780	-0.8930	16.9270
H	-3.8390	0.4520	16.8100
N	-2.2140	2.3170	16.6880
C	-1.1010	3.0940	16.6390
C	-2.9500	7.6070	16.2350
H	-2.5860	8.6290	16.3460
C	-2.1480	6.6790	17.1470
H	-1.3440	7.2410	17.6220

H	-2.8040	6.2670	17.9140
O	-4.3240	7.5450	16.5930
P	-4.7280	7.6260	18.1390
O	-6.0090	8.3560	18.2800
O	-3.5920	8.1640	18.9210
O	-4.9450	6.0810	18.4920
C	-5.5820	5.2270	17.5240
H	-4.8550	4.8090	16.8280
H	-6.3060	5.8290	16.9750
C	-6.3060	4.0950	18.2270
H	-7.0380	3.6070	17.5840
O	-5.3130	3.0990	18.4570
C	-5.0020	2.9580	19.8280
H	-5.2720	1.9610	20.1780
N	-3.5150	2.9690	19.9100
C	-2.8190	4.1380	19.8810
H	-3.3460	5.0770	19.7980
C	-1.4640	4.1480	19.9550
H	-0.9240	5.0950	19.9300
C	-0.8110	2.8780	20.0650
N	0.5090	2.8130	20.1420
H	0.9300	1.8980	20.2190
H	1.0620	3.6580	20.1240
N	-1.5060	1.7330	20.0930
C	-2.8610	1.7440	20.0170
O	-3.5360	0.7080	20.0390
C	-6.8580	4.4200	19.6150
H	-7.1640	5.4600	19.7260
C	-5.6640	4.1410	20.5270
H	-5.3440	5.0690	21.0020
H	-5.9520	3.4220	21.2940
O	-7.9330	3.5620	19.9730
P	-8.3080	3.3900	21.5190
O	-9.7730	3.2280	21.6600
O	-7.7050	4.4930	22.3010
O	-7.5750	2.0130	21.8720
C	-7.5880	0.9480	20.9040
H	-6.7540	1.0370	20.2080
H	-8.5280	1.0100	20.3550
C	-7.5090	-0.3940	21.6070
H	-7.8150	-1.2190	20.9640
O	-6.1200	-0.6160	21.8370
C	-5.7850	-0.5470	23.2080
H	-5.4170	-1.5120	23.5580
N	-4.5890	0.3360	23.2900
C	-4.7130	1.6900	23.2610
H	-5.6910	2.1400	23.1780
C	-3.6230	2.4950	23.3350
H	-3.7430	3.5780	23.3100
C	-2.3480	1.8520	23.4450
N	-1.2420	2.5750	23.5220
H	-0.3640	2.0830	23.5980
H	-1.2910	3.5840	23.5040
N	-2.2370	0.5170	23.4730
C	-3.3400	-0.2700	23.3970
O	-3.2770	-1.5060	23.4190
C	-8.1460	-0.4550	22.9950
H	-9.0050	0.2070	23.1050
C	-7.0160	0.0210	23.9070
H	-7.3020	0.9590	24.3820
H	-6.8260	-0.7300	24.6740
O	-8.5120	-1.7810	23.3530
P	-8.7140	-2.1400	24.8990
O	-9.8040	-3.1330	25.0400
O	-8.8740	-0.8940	25.6810
O	-7.3120	-2.8240	25.2520
C	-6.6960	-3.6930	24.2840
H	-6.0740	-3.1300	23.5890
H	-7.4920	-4.1950	23.7340
C	-5.8430	-4.7320	24.9870
H	-5.6060	-5.5790	24.3430
O	-4.5890	-4.0960	25.2170
C	-4.3590	-3.8430	26.5880
H	-3.4950	-4.4070	26.9380
N	-3.9100	-2.4250	26.6700
C	-4.8060	-1.4030	26.6410
H	-5.8620	-1.6150	26.5580
C	-4.3980	-0.1110	26.7150
H	-5.1320	0.6950	26.6900
C	-2.9880	0.1180	26.8250
N	-2.5180	1.3540	26.9020
H	-1.5180	1.4720	26.9780
H	-3.1500	2.1410	26.8840
N	-2.1140	-0.8960	26.8530
C	-2.5430	-2.1820	26.7770
O	-1.7660	-3.1440	26.7990
C	-6.3230	-5.1570	26.3750
H	-7.4070	-5.1270	26.4860
C	-5.6880	-4.1070	27.2870
H	-6.4710	-3.5160	27.7620
H	-5.0930	-4.6030	28.0540
O	-5.8390	-6.4440	26.7330
H	-6.1150	-6.7540	27.5990
H	7.6200	2.9210	29.1350
O	7.5750	2.0130	28.8280
C	7.5880	0.9480	29.7960

H	6.7540	1.0370	30.4920
H	8.5280	1.0100	30.3450
C	7.5090	-0.3940	29.0930
H	7.8150	-1.2190	29.7360
O	6.1200	-0.6160	28.8630
C	5.7850	-0.5470	27.4920
H	5.4180	-1.5120	27.1410
N	4.5890	0.3360	27.4100
C	4.5200	1.7120	27.4240
H	5.4240	2.2970	27.5050
N	3.3000	2.1860	27.3350
C	2.5060	1.0430	27.2560
C	1.0970	0.9160	27.1430
O	0.2460	1.7990	27.0870
N	0.7070	-0.4300	27.0930
H	-0.2710	-0.6230	27.0110
C	1.5660	-1.5100	27.1460
N	0.9960	-2.7160	27.0830
H	0.0090	-2.9190	27.0130
H	1.6170	-3.5110	27.1300
N	2.8880	-1.3900	27.2520
C	3.2830	-0.0910	27.3010
C	8.1460	-0.4550	27.7050
H	9.0050	0.2070	27.5950
C	7.0160	0.0210	26.7930
H	7.3020	0.9590	26.3180
H	6.8260	-0.7300	26.0260
O	8.5120	-1.7810	27.3470
P	8.7140	-2.1400	25.8010
O	9.8040	-3.1330	25.6600
O	8.8740	-0.8940	25.0190
O	7.3120	-2.8240	25.4480
C	6.6960	-3.6930	26.4160
H	6.0740	-3.1300	27.1110
H	7.4920	-4.1950	26.9660
C	5.8430	-4.7320	25.7130
H	5.6060	-5.5790	26.3570
O	4.5890	-4.0960	25.4830
C	4.3590	-3.8430	24.1120
H	3.4950	-4.4080	23.7610
N	3.9100	-2.4250	24.0300
C	4.6630	-1.2710	24.0440
H	5.7380	-1.3280	24.1250
N	3.9540	-0.1710	23.9550
C	2.6400	-0.6290	23.8760
C	1.4260	0.0960	23.7630
O	1.2570	1.3110	23.7070
N	0.3190	-0.7640	23.7130
H	-0.5860	-0.3460	23.6320
C	0.3790	-2.1430	23.7660
N	-0.7910	-2.7830	23.7030
H	-1.7090	-2.3670	23.6330
H	-0.7570	-3.7910	23.7500
N	1.5190	-2.8220	23.8720
C	2.6030	-2.0030	23.9210
C	6.3230	-5.1570	24.3250
H	7.4070	-5.1270	24.2140
C	5.6880	-4.1070	23.4130
H	6.4710	-3.5160	22.9380
H	5.0930	-4.6030	22.6460
O	5.8390	-6.4440	23.9670
P	5.7920	-6.8540	22.4210
O	6.0900	-8.2970	22.2800
O	6.6540	-5.9390	21.6390
O	4.2550	-6.5820	22.0680
C	3.2460	-6.9240	23.0360
H	3.0730	-6.1030	23.7320
H	3.5950	-7.7980	23.5850
C	1.9460	-7.2630	22.3330
H	1.2560	-7.8090	22.9770
O	1.3050	-6.0110	22.1030
C	1.2680	-5.6710	20.7320
H	0.2370	-5.6200	20.3800
N	1.7380	-4.2600	20.6500
C	3.0250	-3.7690	20.6640
H	3.8610	-4.4480	20.7450
N	3.0980	-2.4630	20.5750
C	1.7660	-2.0610	20.4960
C	1.2100	-0.7600	20.3830
O	1.7870	0.3220	20.3270
N	-0.1910	-0.8060	20.3330
H	-0.6770	0.0640	20.2510
C	-0.9530	-1.9560	20.3860
N	-2.2750	-1.7870	20.3230
H	-2.7730	-0.9110	20.2530
H	-2.8390	-2.6230	20.3700
N	-0.4290	-3.1760	20.4920
C	0.9280	-3.1500	20.5410
C	2.0840	-7.8880	20.9450
H	2.9780	-8.5010	20.8340
C	2.1880	-6.6660	20.0330
H	3.1690	-6.6490	19.5580
H	1.4150	-6.7170	19.2670
O	0.9360	-8.6450	20.5870
P	0.6570	-8.9490	19.0410
O	0.0500	-10.2920	18.9000

O	1.8920	-8.7160	18.2590
O	-0.4270	-7.8260	18.6880
C	-1.4430	-7.5100	19.6560
H	-1.1000	-6.7450	20.3530
H	-1.6740	-8.4230	20.2050
C	-2.6950	-7.0200	18.9530
H	-3.5740	-7.0560	19.5960
O	-2.4770	-5.6300	18.7230
C	-2.3080	-5.3330	17.3520
H	-3.1120	-4.6850	17.0020
N	-1.0980	-4.4680	17.2700
C	0.1510	-5.0050	17.2990
H	0.2760	-6.0740	17.3820
C	1.2540	-4.2170	17.2250
H	2.2470	-4.6650	17.2500
C	1.0350	-2.8050	17.1150
N	2.0660	-1.9770	17.0380
H	1.8700	-0.9890	16.9620
H	3.0100	-2.3360	17.0560
N	-0.1990	-2.2870	17.0870
C	-1.2890	-3.0930	17.1630
O	-2.4450	-2.6510	17.1410
C	-2.9500	-7.6070	17.5650
H	-2.5860	-8.6290	17.4540
C	-2.1480	-6.6790	16.6530
H	-1.3440	-7.2410	16.1780
H	-2.8040	-6.2670	15.8860
O	-4.3240	-7.5450	17.2070
P	-4.7280	-7.6260	15.6610
O	-6.0090	-8.3560	15.5200
O	-3.5920	-8.1640	14.8790
O	-4.9450	-6.0810	15.3080
C	-5.5820	-5.2270	16.2760
H	-4.8550	-4.8090	16.9720
H	-6.3060	-5.8290	16.8250
C	-6.3060	-4.0950	15.5730
H	-7.0380	-3.6070	16.2160
O	-5.3130	-3.0990	15.3430
C	-5.0020	-2.9580	13.9720
H	-5.2720	-1.9620	13.6210
N	-3.5150	-2.9690	13.8900
C	-2.6500	-4.0420	13.9040
H	-3.0370	-5.0470	13.9850
N	-1.3850	-3.7080	13.8150
C	-1.4140	-2.3170	13.7360
C	-0.3490	-1.3860	13.6230
O	0.8580	-1.6000	13.5670
N	-0.8250	-0.0680	13.5730
H	-0.1480	0.6630	13.4920
C	-2.1550	0.3010	13.6260
N	-2.4020	1.6120	13.5630
H	-1.7220	2.3560	13.4930
H	-3.3720	1.8910	13.6100
N	-3.1530	-0.5730	13.7320
C	-2.7090	-1.8560	13.7810
C	-6.8580	-4.4200	14.1850
H	-7.1640	-5.4600	14.0740
C	-5.6640	-4.1410	13.2730
H	-5.3440	-5.0690	12.7980
H	-5.9520	-3.4220	12.5060
O	-7.9330	-3.5620	13.8270
P	-8.3080	-3.3900	12.2810
O	-9.7730	-3.2280	12.1400
O	-7.7050	-4.4930	11.4990
O	-7.5750	-2.0130	11.9280
C	-7.5880	-0.9480	12.8960
H	-6.7540	-1.0370	13.5920
H	-8.5280	-1.0100	13.4450
C	-7.5090	0.3940	12.1930
H	-7.8150	1.2190	12.8360
O	-6.1200	0.6160	11.9630
C	-5.7850	0.5470	10.5920
H	-5.4170	1.5120	10.2420
N	-4.5890	-0.3360	10.5100
C	-4.7130	-1.6900	10.5390
H	-5.6910	-2.1400	10.6220
C	-3.6230	-2.4950	10.4650
H	-3.7430	-3.5780	10.4900
C	-2.3480	-1.8520	10.3550
N	-1.2420	-2.5750	10.2780
H	-0.3640	-2.0830	10.2020
H	-1.2910	-3.5840	10.2960
N	-2.2370	-0.5170	10.3270
C	-3.3400	0.2700	10.4030
O	-3.2770	1.5060	10.3810
C	-8.1460	0.4550	10.8050
H	-9.0050	-0.2070	10.6950
C	-7.0160	-0.0210	9.8930
H	-7.3020	-0.9590	9.4180
H	-6.8260	0.7300	9.1260
O	-8.5120	1.7810	10.4470
P	-8.7140	2.1400	8.9010
O	-9.8040	3.1330	8.7600
O	-8.8740	0.8940	8.1190
O	-7.3120	2.8240	8.5480
C	-6.6960	3.6930	9.5160

H	-6.0740	3.1300	10.2110
H	-7.4920	4.1950	10.0660
C	-5.8430	4.7320	8.8130
H	-5.6060	5.5790	9.4570
O	-4.5890	4.0960	8.5830
C	-4.3590	3.8430	7.2120
H	-3.4950	4.4080	6.8610
N	-3.9100	2.4250	7.1300
C	-4.6630	1.2710	7.1440
H	-5.7380	1.3280	7.2250
N	-3.9540	0.1710	7.0550
C	-2.6400	0.6290	6.9760
C	-1.4260	-0.0960	6.8630
O	-1.2570	-1.3110	6.8070
N	-0.3190	0.7640	6.8130
H	0.5860	0.3460	6.7320
C	-0.3790	2.1430	6.8660
N	0.7910	2.7830	6.8030
H	1.7090	2.3670	6.7330
H	0.7570	3.7910	6.8500
N	-1.5190	2.8220	6.9720
C	-2.6030	2.0030	7.0210
C	-6.3230	5.1570	7.4250
H	-7.4070	5.1270	7.3140
C	-5.6880	4.1070	6.5130
H	-6.4710	3.5160	6.0380
H	-5.0930	4.6030	5.7460
O	-5.8390	6.4440	7.0670
P	-5.7920	6.8540	5.5210
O	-6.0900	8.2970	5.3800
O	-6.6540	5.9390	4.7390
O	-4.2550	6.5820	5.1680
C	-3.2460	6.9240	6.1360
H	-3.0730	6.1030	6.8320
H	-3.5950	7.7980	6.6850
C	-1.9460	7.2630	5.4330
H	-1.2560	7.8090	6.0770
O	-1.3050	6.0110	5.2030
C	-1.2680	5.6710	3.8320
H	-0.2370	5.6200	3.4800
N	-1.7380	4.2600	3.7500
C	-3.0250	3.7690	3.7640
H	-3.8610	4.4480	3.8450
N	-3.0980	2.4630	3.6750
C	-1.7660	2.0610	3.5960
C	-1.2100	0.7600	3.4830
O	-1.7870	-0.3220	3.4270
N	0.1910	0.8060	3.4330
H	0.6770	-0.0640	3.3510
C	0.9530	1.9560	3.4860
N	2.2750	1.7870	3.4230
H	2.7730	0.9110	3.3530
H	2.8390	2.6230	3.4700
N	0.4290	3.1760	3.5920
C	-0.9280	3.1500	3.6410
C	-2.0840	7.8880	4.0450
H	-2.9780	8.5010	3.9340
C	-2.1880	6.6660	3.1330
H	-3.1690	6.6490	2.6580
H	-1.4150	6.7170	2.3670
O	-0.9360	8.6450	3.6870
P	-0.6570	8.9490	2.1410
O	-0.0500	10.2920	2.0000
O	-1.8920	8.7160	1.3590
O	0.4270	7.8260	1.7880
C	1.4430	7.5100	2.7560
H	1.1000	6.7450	3.4530
H	1.6740	8.4230	3.3050
C	2.6950	7.0200	2.0530
H	3.5740	7.0560	2.6960
O	2.4770	5.6300	1.8230
C	2.3080	5.3330	0.4520
H	3.1120	4.6860	0.1010
N	1.0980	4.4680	0.3700
C	-0.2320	4.8270	0.3840
H	-0.5090	5.8680	0.4650
N	-1.0590	3.8140	0.2950
C	-0.2180	2.7050	0.2160
C	-0.5320	1.3260	0.1030
O	-1.6350	0.7900	0.0470
N	0.6280	0.5400	0.0530
H	0.5090	-0.4500	-0.0290
C	1.9210	1.0230	0.1060
N	2.8910	0.1080	0.0430
H	2.7780	-0.8930	-0.0270
H	3.8390	0.4520	0.0900
N	2.2140	2.3170	0.2120
C	1.1010	3.0940	0.2610
C	2.9500	7.6070	0.6650
H	2.5860	8.6290	0.5540
C	2.1480	6.6790	-0.2470
H	1.3440	7.2410	-0.7220
H	2.8040	6.2670	-1.0140
O	4.3240	7.5450	0.3070
H	4.5340	7.9030	-0.5590
Na	5.2048	-8.0672	3.8233

Na	8.5147	-3.3530	7.3796
Na	8.8594	2.2923	10.7596
Na	5.8205	7.0616	14.1396
Na	0.5576	9.1339	17.5196
Na	-4.9172	7.7179	20.8996
Na	**-8.5147**	**3.3530**	**24.2796**
Na	-10.8572	-1.8907	26.5159
Na	10.8572	-1.8907	24.1841
Na	5.8205	-7.0616	19.6604
Na	0.5576	-9.1339	16.2804
Na	-4.9172	-7.7179	12.9004
Na	-8.5147	-3.3530	9.5204
Na	-8.8594	2.2923	6.1404
Na	-5.8205	7.0616	2.7604
Na	-0.5305	9.5858	-0.4433