# Single Molecule Protein Conductance Measurements:

# Novel Methods of Experimental Data Analysis

by

Sohini Mukherjee

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2023 by the
Graduate Supervisory Committee:

Stuart Lindsay, Chair
Thomas Moore
Quan Qing

ARIZONA STATE UNIVERSITY

May 2023

# ABSTRACT

Exploration of long-range conductance in non-redox-active proteins at the single-molecule scale is aided by the development of innovative, tailor-made quantitative data analysis techniques. This thesis details the rationale behind the proposed approaches, the steps taken to design and implement every method, and the validation of the methodologies using appropriate experiments, benchmarks, and rigorous statistical data analysis. The first chapter conducts a thorough literature review, sets the stage for the subsequent investigation, and underscores the importance of the research questions addressed in this thesis. The second chapter describes the solvent effects on the electronic conductance of a series of Consensus Tetratricopeptide Repeat proteins (CTPR) measured with Scanning Tunneling Microscopy (STM). The study reveals a reversible reduction in electronic conductance when water ($H_2O$) is replaced with heavy water ($D_2O$) due to a $\sim$ 6-fold decrease in the carrier diffusion constant as proteins become solvated by $D_2O$. Similar observations are made in a $\sim$7 nm long tryptophan zipper protein, while a phenylalanine zipper protein of comparable length remains unchanged in $D_2O$, highlighting the critical role of aromatic residues in proteins lacking redox cofactors. As an extension to this finding, the third chapter describes the development of a machine-learning model to detect the presence of a protein and identify essential features helping in the detection. For this purpose, a solid-state device was engineered to measure the conductance of CTPR-16 protein wires. This approach addresses the limitations in characterizing the STM gap, enables the collection of stable current vs. time data, and provides a statistical understanding of the electronic transport through a protein. The final chapter investigates real-time changes in conductance in response to protein conformation alterations. A deoxyribonucleic acid (DNA) polymerase $\Phi$29 was chosen for its potential utility as a single-molecule DNA sequencing device. The modified enzyme was bound to electrodes functionalized with streptavidin. $\Phi$29 connected by one biotinylated contact and a second nonspecific contact showed rapid small fluctuations in current when activated. Signals were greatly enhanced with two specific contacts. Features in the distributions of conductance increased by a factor 2 or more over the open-to-closed conformational transition of the polymerase.

*To my grandfather.*

# ACKNOWLEDGEMENTS

Writing a thesis is a journey that can be challenging and rewarding, and I couldn't have made it without the support of many individuals who made a significant contribution to my success. I am deeply grateful for the assistance of those who have supported and guided me on this journey.

First and foremost, I extend my deepest gratitude to my advisor, Professor Stuart Lindsay, for his unwavering support and mentorship throughout my research. His encouragement, constructive criticism, and feedback were invaluable, and I could not have completed this thesis without his guidance. I will always cherish the discussions and brainstorming sessions with him.

I would also like to thank my committee members, Professors Thomas Moore and Quan Qing, for their constructive feedback and invaluable insights that helped shape my research project. I would also take this opportunity to thank Professor Antonia Papandreou-Suppappola for her critical review and helpful suggestions that made a significant contribution to the quality of this thesis.

I am grateful to the faculty and staff of the School of Molecular Sciences and The Biodesign Institute at Arizona State University, who provided me with an excellent academic environment, resources, and support during my studies.

The projects would not have gained fruition without the immense expertise of Dr. Sepideh Afsari, Dr.Bintian Zhang, Dr. Weisi Song, Dr. Brian Ashcroft, Dr. Eathen Ryan, Dr. Nicholas Halloran, and Dr. Hanqing Deng.

I owe a debt of gratitude to my most significant supporter of all time, my mother, Smita Mukherjee; the best sisters in the entire Universe, Sreya Mukherjee and Aritrika Sarkar; my father, Partha Pratim Mukherjee, for their unconditional love, unwavering support and encouragement throughout my life, especially during my academic journey. I especially thank my partner and soul-mate, Gouranga Charan, for not just

believing in me and guiding me during challenging times but also making me a part of his beautiful family, Parbati Charan, Jadunath Charan, and Nittananda Charan. I also thank my uncle, Subir Sarkar, and aunt, Polly Sarkar, for always having my back.

I would like to thank my amazing friends, Sohang, Trisha, Rishav, Pallav, Upasana, Nishit, Abesh, Saborni, Mayukh, Pritha, Subhankar, Ravish, Souvik, Shrestha, Anisha, Akanksha, Joydeep, Sanchari, Raghu, Leeza, Reecheek, Abhishek, Sumanta, and many more who shared this journey with me, providing moral support and encouragement during the highs and lows of the Ph.D. process. Your camaraderie and support have been invaluable, and I am grateful for the memories we shared.

Finally, I would like to remember my brother, Arunava Saha, who left me suddenly, leaving a void that can never be filled. Your memories will keep me going.

In conclusion, this Ph.D. thesis was a collaborative effort, and I am humbled and honored to have had so many incredible people by my side throughout this journey. Their contributions to this thesis and my life are immeasurable, and I am forever grateful.

## TABLE OF CONTENTS

CHAPTER Page

APPENDIX

# LIST OF TABLES

Table Page

# LIST OF FIGURES

Figure Page

# Chapter 1

# INTRODUCTION

This chapter lays the foundation for the research work presented in the later chapters. The first part of the introduction (Section 1.1) focuses on single-molecule conductance measurements. This section includes a description of the Scanning Tunneling Microscopy (STM) setup and provides a historical account of the research work done on single-molecule conductance measurements in proteins that lack redox co-factors. Additionally, this section highlights the contribution of the thesis to the ongoing research in this area. Section 1.2 of the introduction outlines the proteins of interest in the thesis. The section provides an explanation of the structure and function of Consensus Tetratricopeptide Repeat (CTPR) Proteins, Phenylalanine, and Tryptophan Zipper Proteins, which are the core of Chapters 2 and 3. This section also describes the Φ29 DNA Polymerase, which is studied extensively in Chapter 4. Section 1.3 of this chapter explores the multifaceted approaches used for quantitative data interpretation in the thesis. These approaches include Asymmetric Least Squares for baseline drift correction in STM data ( used in Chapter 2), Matrix Profiling for pattern recognition in temporal electrical current signature changes during DNA - Φ29 DNA Polymerase interaction (used in Chapter 4), and Deep Learning-Aided Binary Classification for binary classification to detect the presence of a specific protein and identify crucial features necessary for the detection process (used in Chapter 3). Gaussian mixture modeling is used for estimating the underlying distribution parameters of single molecule conductance data for CTPR proteins and Zipper proteins in Chapter 2.

## 1.1 Single Molecule Conductance Measurements

### *1.1.1 Scanning Tunneling Microscopy*

The first single-molecule study, conducted in 1961, measured the activity of single molecules of beta-D-galactosidase in microdroplets and on a fluorogenic substrate [1, 2]. Since then, various single-molecule techniques have been developed across disciplines, from physics to biology [3]. Examples of single-molecule techniques include scanning tunneling microscopy (STM), atomic force microscopy (AFM), ion traps, atom traps, confocal microscopy with fluorescence, and optical tweezers [1]. While some methods require extreme conditions, such as low temperatures or ultra-high vacuums, others can be performed in liquid at room temperature. These techniques can be applied to a wide range of molecules, from single atoms to complex living cells [4–7].

Electrons serve as excellent probes for single-molecule measurements, as their tunneling behavior enables measurements at nanometer scales with angstrom sensitivity. Understanding charge transport at the single-molecule level provides chemical information and can be applied to molecular electronics and sensor applications that rely on electrical detection of molecular binding events [4, 8–10].

Electron-based single-molecule measurement techniques, such as Scanning Tunneling Microscopy (STM), offer direct and label-free detections. These techniques, including nano-gap methods, can measure molecular conductance, vibrational energy levels, electronic polarizability, and spin states [11, 12]. Quantum tunneling allows electrons to be transported across a nanometer gap, such as a molecule or insulating layer between electrodes. The sensitivity of tunneling current to gap size and medium suggests that solution environments provide opportunities to study chemical reactions and biopolymer sequencing [13].

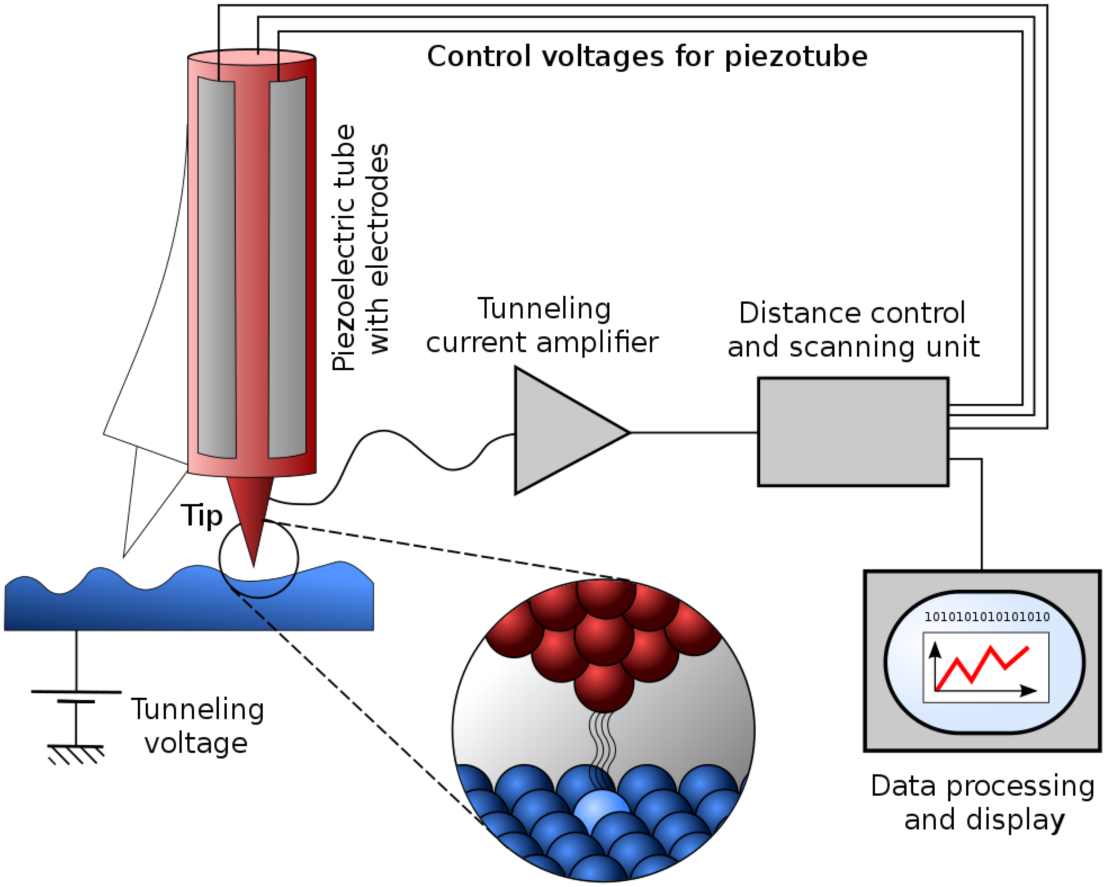

Control voltages for piezotube
Piezoelectric tube with electrodes
Tunneling current amplifier
Distance control and scanning unit
Tip
Tunneling voltage
1010101010101010
Data processing and display

Figure 1.1: Schematics of Scanning Tunneling Microscopy (STM). STM is based in tunneling phenomena between metal tip and metal sample. At a few nanometer gap, a few pico-ampere of tunneling current is induced. An amplifier is required to read very small tunneling current, and the small gap is maintained by feedback controller and piezoelectric tube [14].

An STM setup (Fig. 1.1) consists of a sharp metal tip, typically made of materials like tungsten, platinum-iridium, gold or palladium, brought extremely close to the sample surface. A voltage bias applied between the tip and the sample results in a measurable tunneling current, extremely sensitive to the tip-sample distance. Piezoelectric materials control the tip's position with atomic-level accuracy. A feedback loop maintains either a constant current or distance between the tip and sample surface, generating an image of the surface topography or electronic properties at an atomic scale with high resolution and sensitivity. With its ability to image individual molecules absorbed on a substrate and manipulate single molecules or atoms, STM can provide chemically-sensitive measurements by modifying the tip with a molecule [4, 10, 15, 16].

The advent of scanning tunneling microscopy in the 1980s revolutionized molecular analysis. Small organic molecules could be temporarily trapped between two metal electrodes with sub-nanometer separation, with tunneling currents between the electrodes revealing the molecular signature of the analyte.Significant advancements have been made in using recognition tunneling for single-molecule amino acid and protein analysis. The method includes covalent modification of electrodes with adaptor molecules that form transient yet well-defined connections with target molecules leading to rapid and fluctuating tunnel current signals.[17],[18] These signals are further processed using machine learning algorithms, enabling the differentiation of individual amino acids and small peptides. More advancements are discussed in the next section.

### *1.1.2 Long-range Conductance in Non-Redox-Active Proteins*

It was widely assumed that proteins acted as insulators [19, 20], with reports of metallic conduction in bacterial wires [21, 22], and long-range transport in protein multilayers [23] being seen as exceptions. However, recent research has suggested that many proteins evolve towards a quantum critical state [24], motivating further investigation into their conductive properties. It was observed that electrical charges are transmittable through proteins if a protein bridges the electrodes via chemical bond formation or ligand binding. The next part discusses the evolution of single-molecule protein conductance measurements and plausible electron transport mechanisms. Single - molecule conductance measurements were conducted using Scanning Tunneling

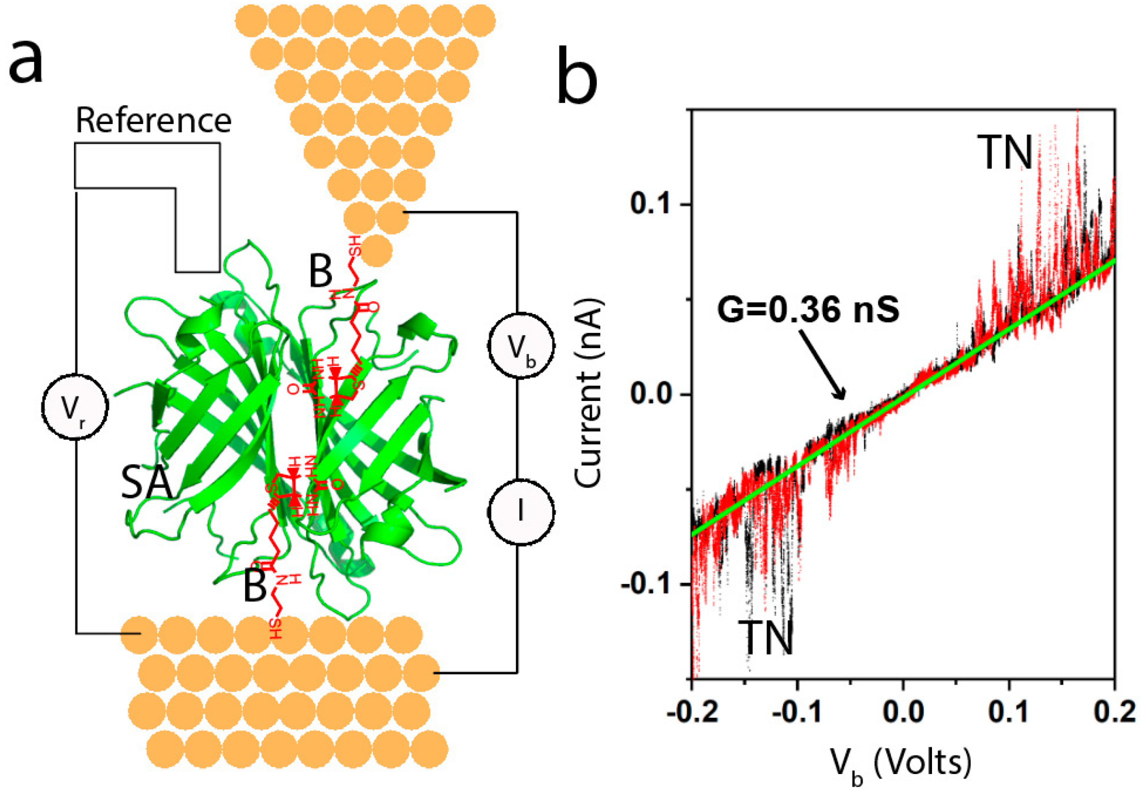


Figure 1.2: Example of the Experimental Set-up for Electrochemical STM Studies of Single-molecule Conductance. (a) Electrochemical STM. The electrodes are functionalized with thiolated biotin molecules (B—red on the diagram) shown here trapping a streptavidin protein (SA—green). (b) Example of a current-voltage curve obtained from a trapped protein by sweeping V and measuring I. The black data points are from the sweep up and the red data points are from the sweep down. Here, TN stands for telegraph noise. Figure from ref. [25]

Microscopy explained in Section 1.1.1 with electrodes submerged in electrolyte and under electrochemical potential control, ensuring electrode potentials remained outside the region where Faradaic currents are generated. Research has focused on electrochemically inert proteins to avoid potential redox cofactor - mediated transport phenomena arising via rapid reduction and oxidation of redox - active sites. Also, treating the electrode surfaces with specific ligands for the target protein helped in avoiding non - specific adsorption and protein denaturation. [26] Nanosiemens conductance over approximately 10 nm distance was observed in a large protein (integrin) when it was bound to one of the two electrodes by a specific bond.[27] A non-binding mutant integrin produced no signal. A systematic study was conducted on five proteins (three types of antibody, a Fab fragment, and streptavidin),[28] employing specific ligands for each target protein. To understand the role of non-specific contacts, bare electrodes and electrodes functionalized with mercaptoethanol (hydrophilic) were also examined.[28] These measurements were performed using an electrochemical scanning tunneling microscope (STM) (Fig. 1.2 (a)), with electrodes functionalized with the appropriate ligand and maintained under potential control. The process was repeated at multiple points on the substrate, collecting about 1,000 current-voltage (IV) curves in a single experiment (Fig. 1.2 (b)). These curves are acquired as a fixed gap traps the target molecule, eliminating the stresses and strains associated with break-junction measurements. Many such curves are used to compile distributions of single-molecule conductance. Most of these curves were reproducible and linear (its slope yields the conductance for the particular contact geometry), with telegraph noise (TN) observed in all proteins studied at all gap distances, usually at biases above 100 mV. TN reflects two or more discrete levels to which the current jumps at a constant bias, usually due to the electric field-induced fluctuations at the contact point.[27]

Multiple single - molecule conductance measurements, derived from the slope of current - voltage (IV) curves, produce distributions representing the variety of contact geometries. Bivalent antibodies (an Anti - DNP IgE ,Anti - Ebola IgG and Anti - HIV IgG), each of which presents two binding sites connected by one specific contact and one non - specific contact produced a log - normal distribution (Fig. 1.3 (a)), with a peak value between 0.2 and 0.4 nS whereas a bimodal distribution (Fig. 1.3 (b)) was observed for antibodies capable of binding specifically to both the electrodes, with a second peak at about 10 times the conductance of the first. In the case of a Fab

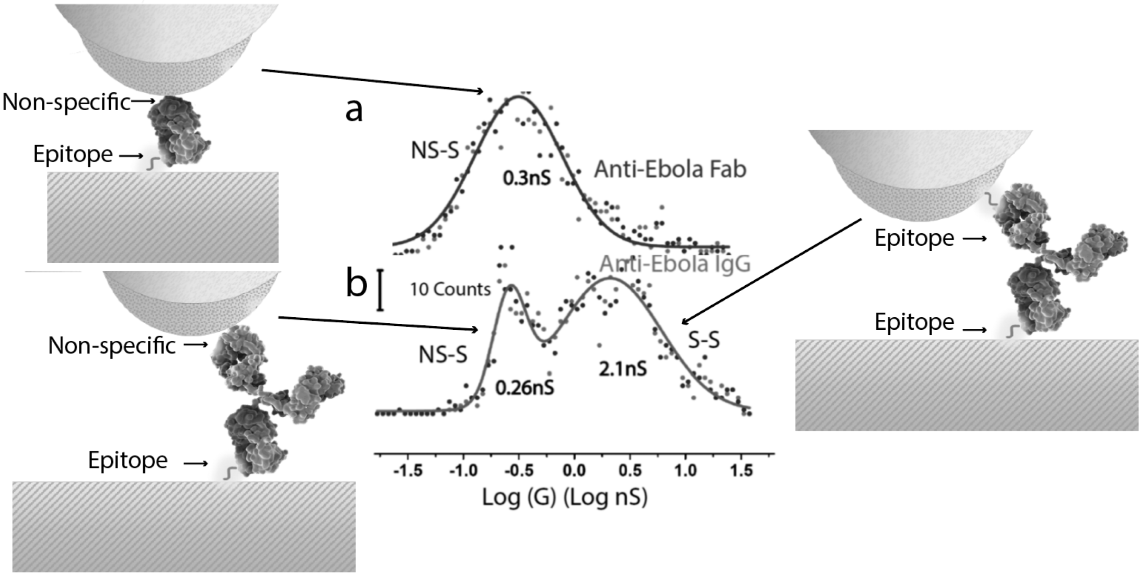


Figure 1.3: Conductance Distributions for an Antibody Fab Fragment and a Full IgG Antibody. (a) an antibody Fab fragment (one specific bond as illustrated on top left) and (b) the full antibody in contact with electrodes functionalized with an epitope for the antibody. The high conductance peak comes from two specific contacts (illustrated on the right) whereas one specific and one non - specific contact gives rise to a smaller conductance peak (as illustrated on the left). Figure from ref. [25]

fragment (half - antibody) from one of the antibodies, the distribution shifted from bimodal (Fig. 1.3 (b)) to a single - peak log - normal distribution (Fig. 1.3 (a)). The results indicated that the current path must enter one Fab fragment of the antibody and exit the other.[28]When tethered by two specific contact points, the insensitivity of the peak position to the gap size was also observed, verifying the existence of the conductance path through the protein. In the case of one non - specific contact (where contact could be made at virtually any point on the antibody surface) and one specific contact, the lower peak showed a statistically significant decrease in conductance with distance. [29] Another study observed that the conductance of proteins in tunnel gaps highly depended on the metal(s) used as contacts. Rest-potential measurements were performed to calibrate the electron injection potentials for specific combinations of metals and surface functionalizations. Fig. 1.4 displays the results for three different

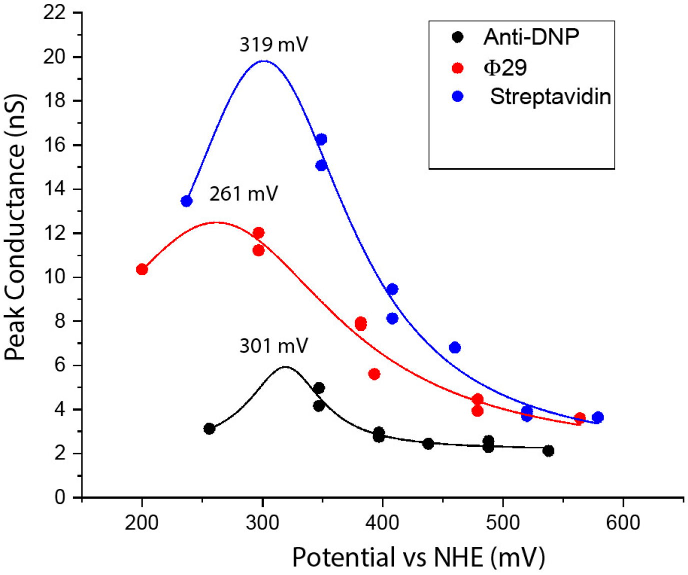


Figure 1.4: Conduction Resonances in Three Proteins, (denoted in the legend), showing peak conductance versus the electron injection potential, calculated from rest potential measurements. [30] Solid lines are fits to Lorentzian functions based on a model of resonant injection with the peak values indicated. Figure from ref. [25]

proteins using combinations of gold, palladium, and platinum electrodes. All three proteins examined exhibit a resonance peak at approximately +300 mV on the normal hydrogen electrode (NHE) scale ( +300 mV NHE is about -4.9 eV with respect to the vacuum [31]). The presence of a resonance peak indicates that electrons are being injected directly into molecular states, with the middle of the band of states located at approximately 300 mV on the NHE scale. The oxidation potential of amino acids determines the energy of their stably ionized states. Tyrosine and tryptophan, the most readily oxidized residues, have redox potentials around +1 V NHE. The observed transmission peaks are 0.7 V away from these redox potentials. There has been a lack of systematic studies of the length dependence of electron or hole transport through intact proteins over a wide range of lengths, except the study done with macroscopic amyloid crystals [34]. Temperature-independent conductance in some protein layers [35–40] has been interpreted as evidence of coherent tunneling transport [41, 42].

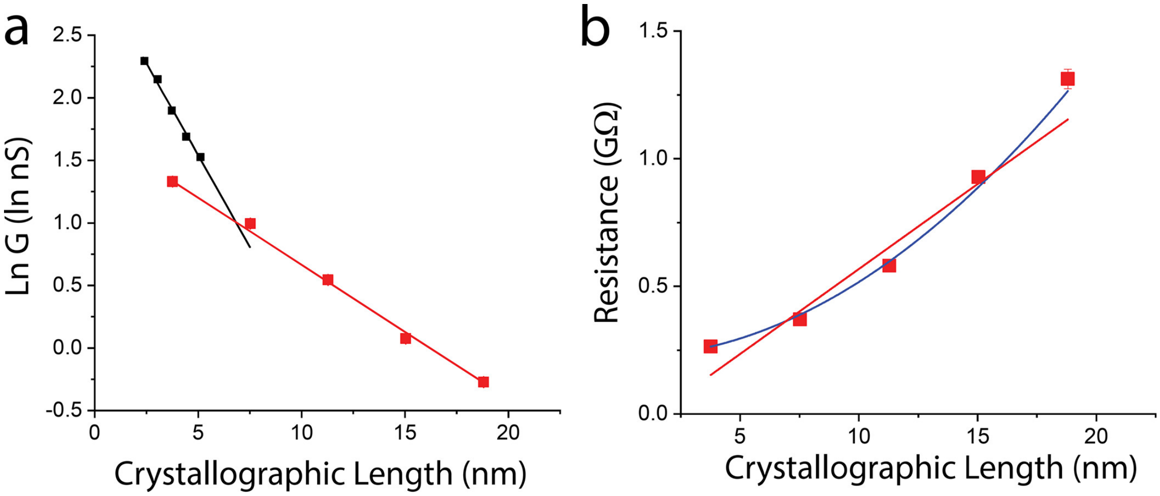


Figure 1.5: Studies of the Length Dependence of Electron or Hole Transport Through CTPR Proteins over a Range of Lengths.(a) Conductance decay fitted to an exponential with a decay constant $1/\lambda = 0.107\acute{s}0.003nm\_1$ (red points in (a), $R^2 = 0.998$).The black data points (data from [32]) are for the region of hopping transport for the organic molecular wire Oligo(p-phenylene ethynylene),showing the relative enhancement offered by protein wires for distances greater than $\sim 6$ nm. (b) The increase of resistance with length departs significantly from the linear behavior usually expected for hopping (red line, $R^2 = 0.925$) but is well-fitted by a square-law dependence (blue line, $R^2 = 0.994$). Error bars are approximately equal to symbol sizes. Figure from ref. [33]

Conductance resonances in redox-active proteins as a function of surface potential [43–47] have been attributed to sequential incoherent hops via a relaxed redox state, [48, 49] or coherent tunneling [41]. To investigate the performance of proteins as molecular wires and settle the debate about whether transport across proteins is dominated by tunneling or hopping, the authors explored the conductance of a series of consensus tetratricopeptide repeat (CTPR) proteins, which are linear structures ranging from 4 to 20 nm in length [33, 50]. The results in Fig. 1.5 showed that the decay of current with distance is slow, allowing these protein wires to outperform the oligo(phenylene ethynylene) (OPE) wires (one of the family “Tour wires” widely used in molecular electronics [32]) in the long-range hopping limit for distances over 6 nm . The long-range transport was found to be dominated by a field-free random diffusion process, not quantum-coherent [51]. In this study, charge injection from noble-metal electrodes was found to resonate with slightly relaxed electronic states of the protein, likely associated with tyrosine and tryptophan residues. This leads to hopping transport. Under a weak driving force, the decay of current exhibits a square law dependence on length. The long diffusion length and the presence of states that deviate significantly from equilibrium oxidation energies suggest hole transport with substantially reduced reorganization energy.

### *1.1.3 Contribution*

This thesis delves into the development of innovative, tailor-made quantitative data analysis techniques for exploring long-range conductance in non-redox-active proteins at the single-molecule scale. The first chapter conducts a thorough literature review and background study, laying a robust foundation for the research. It presents a detailed overview of relevant theories, methodologies, and applications, emphasizing state-of-the-art techniques and identifying various data analysis methods. By critically examining the existing literature, this chapter sets the stage for the subsequent investigation and underscores the importance and novelty of the research questions addressed in this thesis.

In the second chapter, the electronic conductance of a series of linear proteins (Consensus Tetratricopeptide Repeat, CTPR proteins) is measured in both $H_2O$ and $D_2O$, examining the effects of the solvent environment on proteins at the nanoscale. The study reveals a reversible reduction in electronic conductance when $H_2O$is replaced with $D_2O$ due to a ∼6-fold decrease in the carrier diffusion constant as proteins become solvated by $D_2O$. This change in conductance is within a factor of ∼2.5 of the reported values for bacterial wires when normalized for length. Similar observations are made in a ∼7 nm long tryptophan zipper protein, while a phenylalanine zipper protein of comparable length remains unchanged in $D_2O$. This highlights the critical role of aromatic residues in proteins lacking redox cofactors.

The third chapter extends the work from the second chapter by developing a machine-learning model to detect the presence of a protein and identify essential features for the detection process. For this purpose, a solid-state device was engineered, which consists of two bimetallic layers separated by a  6 nm thick insulating layer deposited using plasma-enhanced atomic layer deposition (PEALD) to study the conductance of CTPR16 protein wires. This approach addresses the limitations in characterizing the STM gap and enables the collection of stable current vs. time data with an improved signal-to-noise ratio. This analysis provides a preliminary understanding of electronic current signature changes when a protein bridges the gap in the solid-state device.

The final chapter investigates changes in conductance in response to protein conformation alterations. A study demonstrated that biotin binding to streptavidin significantly impacts its conductance [28]. In multivalent proteins like streptavidin, two binding sites can be used for electrical connections, leaving two sites open for sensing binding events. This chapter focuses on real-time enzyme activity monitoring. With low noise levels in the IV curves below 100 mV, it becomes possible to record the ”noise” generated by an enzyme performing its function by biasing the enzyme with a voltage below 100 mV. However, for enzymes with a single active site, it is necessary to engineer electrical contacts that do not interfere with the protein's function. DNA polymerase Φ29 was chosen for its potential utility as a single-molecule DNA sequencing device.

In summary, each chapter details the rationale behind the proposed approach, the steps taken to design and implement the method, and the validation of the methodology using appropriate experiments, benchmarks, and statistical data analysis.

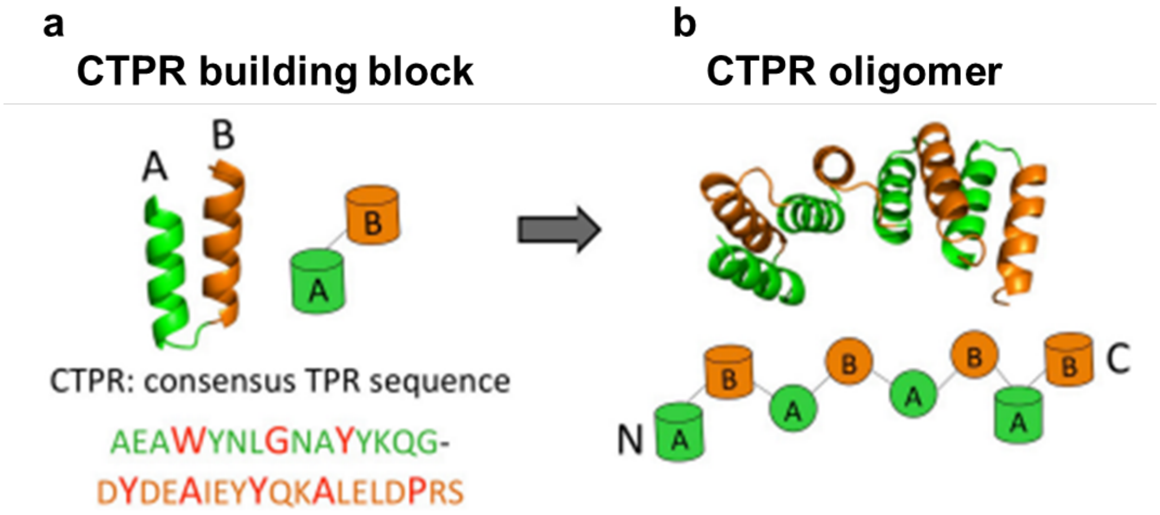


Figure 1.6: CTPR Repeat as a Building Block for Repeat Proteins. (a) The CTPR repeat unit structure is illustrated with helix A in green and helix B in orange. To the right, a schematic representation of the CTPR building block employs the same color scheme. The consensus sequence is displayed below, with conserved amino acids highlighted in red. (b) The crystal structure of a repeat protein containing four CTPR repeats uses green for the A helices and orange for the B helices. Below, a schematic representation of the CTPR packing is shown, extending from the N-terminal to the C-terminal. [52]

### *1.2.1 CTPR Proteins*

Repeat proteins, including the tetratricopeptide repeat (TPR) family, are defined by tandem arrays of a small structural motif, which vary in length (18-47 amino acids) and structure (alpha, beta, or alpha/beta) based on the specific protein family. In order to develop new TPR proteins that encapsulate the sequence-structure relationship of the TPR fold, the Regan Laboratory designed a consensus TPR (CTPR) sequence (Fig. 1.6) by analyzing the statistical properties of natural TPRs [53]. CTPR proteins represent a standardized 34 amino acid helix-turn-helix repeat module that can be combined in tandem to create proteins with various repeat numbers, from 2 to

20 (CTPR2 to CTPR20). CTPR proteins exhibit superhelical structures, with eight repeats forming one complete turn of the superhelix. This unique feature renders CTPR proteins helpful in exploring the structural and functional properties of repeat proteins [54]. A single molecule conductance study of a series of CTPR proteins ranging from 4 nm to 12 nm is described in Chapter 2 and Chapter 3.

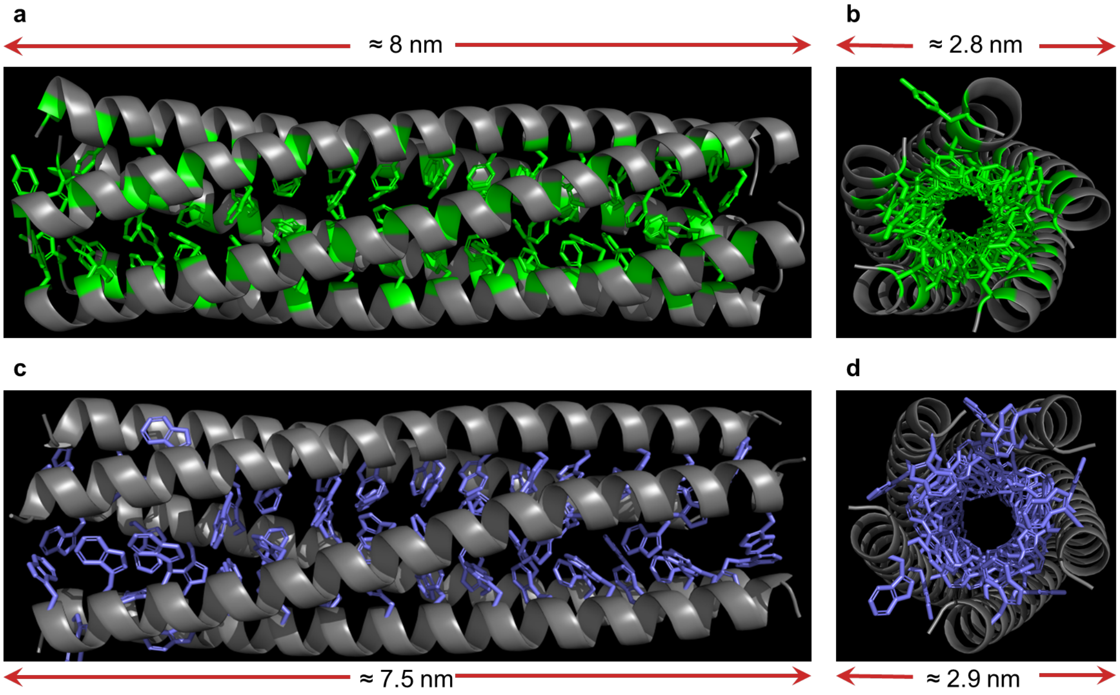


Figure 1.7: Length and Overall Diameter of Zipper Proteins. Phenylalanine zipper showing the published PDB structure (PDB 2GUV). (a) Length, (b) Overall Diameter. Tryptophan zipper showing the published PDB structure (PDB 1T8Z). (c) Length, (d) Overall Diameter.

### *1.2.2 Phenylalanine and Tryptophan Zipper Proteins*

Coiled coils are composed of two to five $\alpha$-helices that entwine around each other in a left-handed superhelical twist. Their structures are primarily dictated by a recurring seven-residue (heptad) sequence labeled as a-b-c-d-e-f-g. Typically, the a and d positions are filled by aliphatic side chains like Leu, Ile, Val, and Ala, while polar residues are found in other positions. The a and d residues allow the $\alpha$-helical side chains to engage through a "knobs-into-holes" pattern, where one helix's side chains (knobs) fit into the spaces (holes) between four side chains of the neighboring helix. As a result, symmetry-related a and d residues create side-by-side interactions, forming interconnected hydrophobic seams throughout the coiled-coil structure's core [55, 56]. Scientists engineered a "Trp-zipper" [56] protein with Trp residues at all 14 a and d positions, discovering that the protein forms a stable alpha-helical pentamer in water at physiological pH. Similarly, they engineered a "Phe-zipper" protein [55] with phenylalanine residues at all 14 hydrophobic a and d positions, which also forms an alpha-helical pentamer. Fig. 1.7 shows that the superhelices create a cylinder (with 5-fold symmetry axis) with an overall diameter of approximately 2.9 nm and 2.9 nm with a length of around 7 nm and 8 nm for the Trp-zipper and the Phe-zipper, respectively. A solvent-based single molecule conductance study of the zipper proteins is described in Chapter 2.

### *1.2.3* Φ29 *DNA Polymerase*

Φ29 DNA polymerase, derived from the Bacillus subtilis bacteriophage Φ29, is a highly processive and accurate enzyme with properties ideal for single-molecule DNA sequencing applications. These properties include high processivity of up to several hundred kilobases, maximum synthesis rates of around 100 bases/s, and a low error rate of approximately 1 in $10^5$ nucleotides incorporated. The enzyme is also capable of strand displacement DNA synthesis, allowing the use of double-stranded DNA as templates. [58, 59] Structurally, as shown in Fig. 1.8, Φ29 DNA polymerase possesses terminal protein regions (TPR1 and TPR2) that are associated with binding terminal proteins. The downstream template DNA moves through a tunnel before reaching the polymerase active site, establishing a structural basis for strand displacement

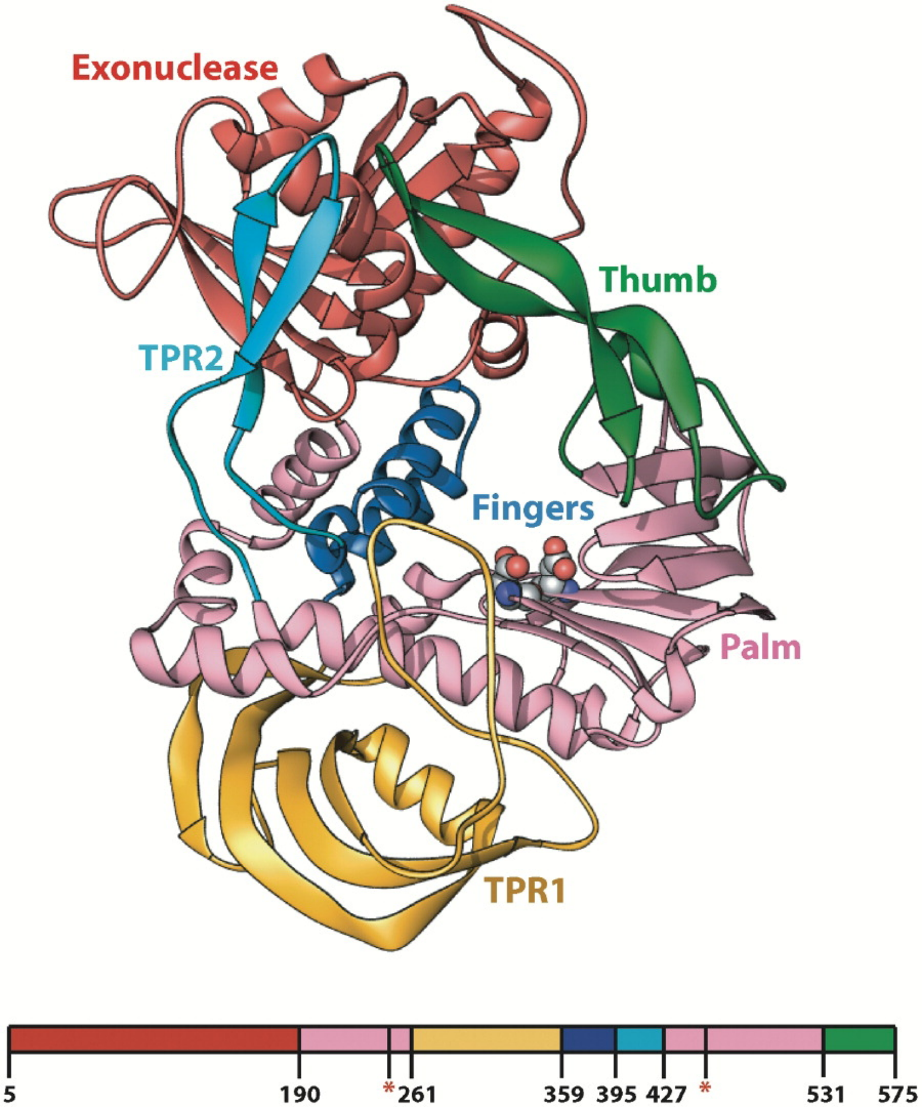


Figure 1.8: Ribbon Representation of the Domain Organization of $\Phi$29 DNA Polymerase. The exonuclease domain is shown in red, the palm in pink, Terminal Protein Regions, TPR1 in gold,TPR2 in cyan, the fingers in blue, and the thumb in green. D249 and D458, which provide the catalytic carboxylates of the polymerase active site, are shown using space-filling spheres.[57]

and processivity. Additionally, the polymerase features a unique thumb, which may function as a clamp to enhance processivity when combined with the polymerase palm subdomain and TPR1 and TPR2 sequences [57]. A carefully engineered $\Phi$29 DNA polymerase enzyme for single molecule conductance measurements is presented in Chapter 4.

## 1.3 Multifaceted Approaches to Quantitative Data Interpretation

### *1.3.1 Asymmetric Least Squares for Baseline Correction*

Asymmetric least squares (ALS) smoothing is a numerical optimization technique used for various applications, including smoothing and baseline correction in signal processing, particularly for spectroscopic data. It involves fitting a smooth curve to the baseline of a spectrum or chromatogram while minimizing the residual error between the curve and the original data. The approach is asymmetric in that it places more weight on the positive residuals (peaks) than on the negative residuals (troughs), which helps to preserve the shape of the peaks while removing the baseline drift, allowing for a more accurate data analysis.

The ALS method is based on the principle of least squares, which aims to minimize the squared differences between the observed data and the fitted model. However, in the asymmetric least squares method, the differences are weighted asymmetrically, meaning that positive and negative deviations from the baseline are treated differently. This is particularly useful when dealing with data where the signal and the baseline have different properties, and more emphasis is needed on either positive or negative deviations. ALS involves a smoothness parameter ($\lambda$) which controls the degree of smoothness of the resulting baseline, with larger values resulting in smoother baselines. Another parameter, the asymmetry parameter (p), determines the balance between the weights assigned to positive and negative deviations, typically ranging from 0 to 1. A value of 0.5 indicates equal weighting, while values closer to 0 or 1 emphasize negative or positive deviations, respectively. At first, the baseline estimate is initialized. This is followed by computing weighted differences between observed

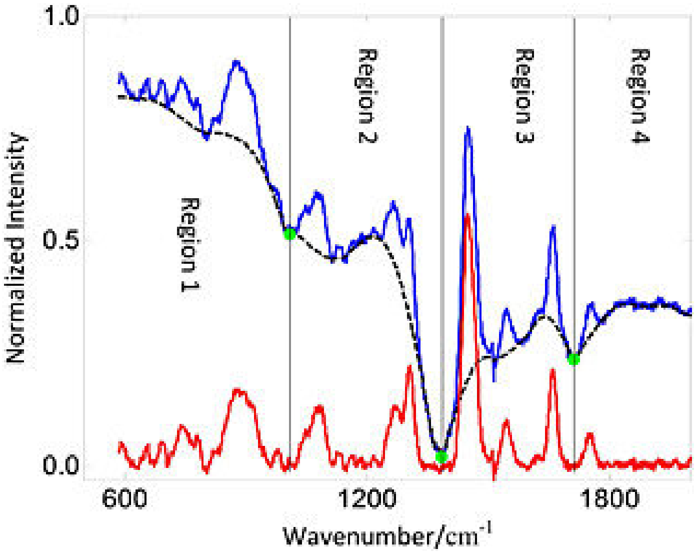


Figure 1.9: Piecewise Asymmetric Least Squares (ALS) for Baseline Correction. Blue line is the spectrum, the black dashed line is the piecewise ALS fit and the red spectrum is the baseline-corrected Raman spectrum obtained by subtracting the piecewise ALS fit from the spectrum. [60]

data and baseline estimates depending on the value of p. Furthermore, the baseline estimate is updated repeatedly to minimize the sum of the weighted square differences, subject to the $\lambda$ value, and achieve convergence. Once the baseline estimate has converged, it can be subtracted from the original data to obtain the corrected signal, which can then be further analyzed or processed as required. An example of ALS-aided baseline correction for a Raman spectrum is shown in Fig. 1.9. The ALS method is flexible and can be adapted to different data types and applications, making it a popular choice for baseline correction and smoothing tasks in various fields. Chapter 4 discusses the employment of this method to the single molecule electronic signatures of proteins.

### *1.3.2 Matrix Profiling for Motif Discovery in Time Series Data*

Introduced in 2016 by Eamonn Keogh at the University of California Riverside and Abdullah Mueen at the University of New Mexico, the Matrix Profile is a versa-

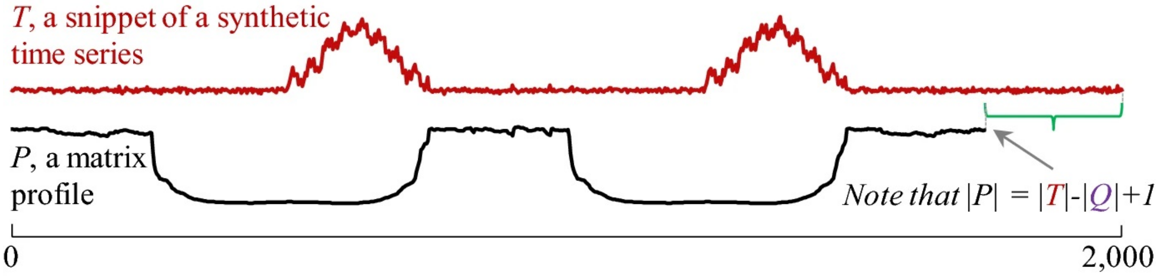


Figure 1.10: A Time Series T, and Its Self-join Matrix Profile P. [61]

tile data structure for time series analysis.[61] Its advantages include being domain-agnostic, fast, and offering exact solutions (or approximate solutions when desired) with only one required parameter. It can identify repeating patterns, also known as motifs, in time series data. It involves computing a matrix that stores the distances between all pairs of sub-sequences in a time series and then using it to find similar motifs in the data.

The Matrix Profile comprises two main components: a distance profile and a profile index. The distance profile is a vector of minimum Z-Normalized Euclidean Distances. In addition, the profile index contains the index of the first nearest-neighbor or the location of the most similar subsequence.

The distance matrix is computed using a sliding window technique that iteratively extracts sub-sequences of fixed length from the time series data and computes the Euclidean distance between each pair of sub-sequences. The distance calculations outlined above occur T-Q + 1 times, where T represents the length of the time series and Q is the window size. Since the subsequences are extracted from the time series itself, an exclusion zone is necessary to prevent trivial matches, such as a segment matching itself or a segment very close to itself. The exclusion zone spans half of the window size (Q) before and after the current window index. The matrix profile is computed by searching for the nearest neighbor of each sub-sequence in the distance

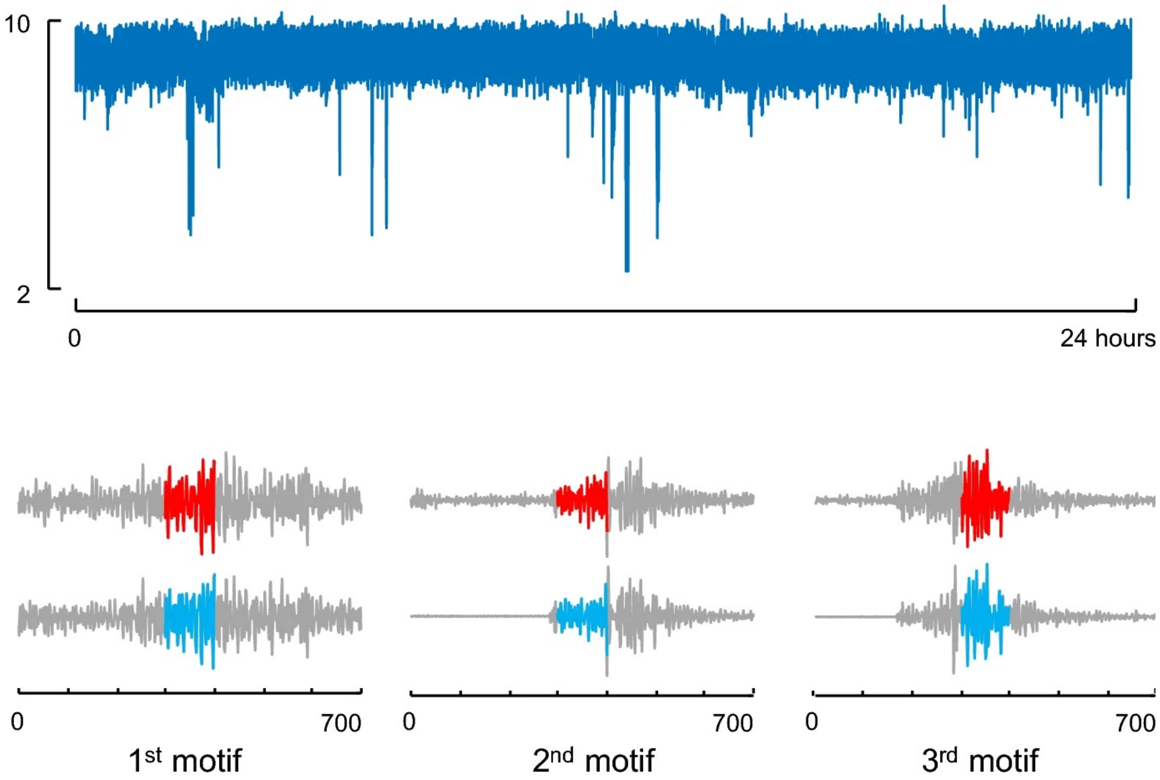


Figure 1.11: Motif Discovery in a given Time Series. (Top) A time-series T. (Bottom) The top 3 motifs extracted from the resulting matrix profile.[61]

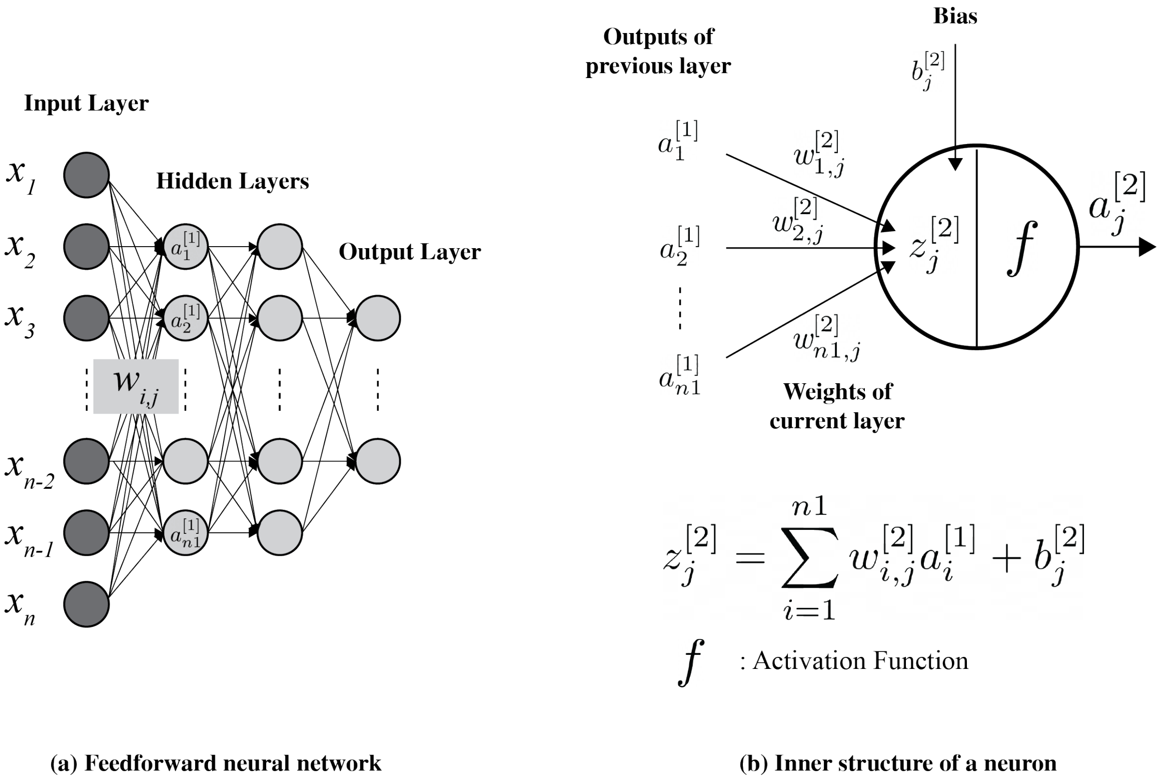
matrix and storing the distance and index of the nearest neighbor in the matrix profile (Fig. 1.10) Finally, the matrix profile can be used to find similar motifs in the time series data by identifying the locations where the matrix profile has low values. These locations correspond to sub-sequences that are similar and can be considered motifs.(Fig. 1.11)

The identified motifs can be characterized by various properties, such as their length, amplitude, frequency, and phase. This analysis can provide insights into the underlying patterns and relationships in the time series data. The utility of this method will be elaborated in Chapter 4.

### *1.3.3 Deep Learning-Aided Binary Classification*

Deep learning is a subfield of machine learning that aims to imitate how humans learn. This is achieved through the use of deep learning algorithms that can learn to extract meaningful features from data, leading to breakthroughs in fields such as image recognition, speech analysis, and natural language processing. One such example of this is the Open-AI ChatGPT, a large language model designed for various downstream NLP tasks, such as text generation/classification and question answering, to name a few. The term ”deep” in deep learning refers to the multiple layers used in the algorithms to extract task-specific high-level features. These algorithms are grouped into two broad categories: supervised and unsupervised. Supervised learning involves using labeled data to train the model, while unsupervised learning utilizes unlabeled data to find patterns and structures within the data. Both these approaches have been successful in various applications and continue to be a significant area of research in machine learning. This thesis adopts a deep learning-based model called the feedforward neural network (FFNN) to perform binary classification. Further details about the objective and the particular architecture of the FFNN are presented in Chapter 3. The following section provides a detailed overview of how an FFNN can be utilized for binary classification. The primary objective of binary classification is to predict one of two classes, represented as 0 or 1, for any given input. In an FFNN, the input layer takes the feature values of the input and passes these values through one or more hidden layers, each consisting of nodes or artificial neurons, to the output layer. For binary classification, the output layer is typically a single node, which returns a value between 0.0 and 1.0, representing the probability that the input belongs to the positive class (1). The details of the FFNN architecture are presented next.

- **Network architecture:** The FFNN consists of an input layer, one or more hidden



(a) Feedforward neural network

(b) Inner structure of a neuron

Figure 1.12: An Example of a One-hidden-layer Neural Network Architecture for Binary Classification. The input layer, activation functions, weights of the hidden layer, and the output class are presented here.

layers, and an output layer. A neural network has two primary components: (i) the weights and (ii) the biases. These parameters determine the network's output for a given input. The weights and biases are learned during training by minimizing the difference between the network's predicted output and the true output for a given input. **Weights** ($w$) are the parameters connecting the neurons in one layer to the neurons in the next layer. Each weight represents the strength of the connection between two neurons. The primary objective of neural network training is to adjust the weights to minimize the difference between the predicted output and the true output for a given input. This process of adjusting the weights is done using an optimization algorithm, such as stochastic gradient descent, explained later under Backpropagation. **Biases** ($b$)

are the parameters that are added to the weighted sum of inputs to each neuron in a layer. Biases allow the network to shift the activation function to the left or right, which can be helpful in modeling complex functions. During training, the biases are adjusted to minimize the difference between the predicted output and the true output for a given input. Together, the weights and biases determine the behavior of the neural network. By adjusting these parameters during the training process, the network can learn to predict the output for a given input accurately. Let $n_l$ denote the number of neurons in layer $l$, $w_{ij}^{(l)}$ the weight connecting neuron $i$ in layer $l-1$ to neuron $j$ in layer $l$, and $b_j^{(l)}$ the bias of neuron $j$ in layer $l$.

- **Activation functions:** The role of the activation functions ($f$) is to introduce non-linearity in the network. Some of the common activation functions utilized in the literature are the sigmoid, Rectified Linear Unit (ReLU) [62], and hyperbolic tangent (tanh), to name a few. For example, the sigmoid function is defined as:

  $$\sigma(x) = \frac{1}{1+e^{-x}}. \tag{1.1}$$

  In general, for binary classification, the output layer has sigmoid function, which reduces the output to a value from 0.0 to 1.0 representing a probability. The hidden layers are usually designed with ReLU activation function for better convergence during training.

The supervised learning process can then be broken down into the following steps:

- **Forward propagation:** The input is passed through the network to compute the output. The weighted sum of the inputs, also called the pre-activation value, for each neuron is calculated as:

$$z_j^{(l)} = \sum_{i=1}^{n_l} w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)}, \tag{1.2}$$

where $a_i^{(l-1)}$ is the activation of neuron $i$ in layer $l-1$, $w_{ij}^{(l)}$ is the weight connecting neuron $i$ in layer $l-1$ to neuron $j$ in layer $l$, and $b_j^{(l)}$ is the bias of neuron $j$ in layer $l$. Then, the activation of neuron $j$ in layer $l$ can be computed as:

$$a_j^{(l)} = f(z_j^{(l)}), \tag{1.3}$$

where $f$ is an user-defined activation function. This process is repeated for all layers until the output layer is reached.

- **Loss function:** A loss function is utilized to measure the difference between the predicted output and the actual output. For binary classification, the binary cross-entropy loss is commonly used:

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{1.4}$$

where $y_i$ is the true label and $\hat{y}_i$ is the predicted probability for the $i$-th sample, and $N$ is the total number of labeled samples in the dataset.

- **Backpropagation:** In this step, the gradients of the loss function are computed with respect to the weights and biases. We start with the output layer:

$$\frac{\partial L}{\partial z_j^{(L)}} = \frac{\partial L}{\partial a_j^{(L)}} \cdot \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}, \tag{1.5}$$

$$\frac{\partial L}{\partial a_j^{(L)}} = \frac{\partial L}{\partial \hat{y}_j}. \tag{1.6}$$

For the sigmoid activation function, we have the following:

$$\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} = \sigma(z_j^{(L)})(1 - \sigma(z_j^{(L)})) = a_j^{(L)}(1 - a_j^{(L)}) \tag{1.7}$$

Now,the gradients for the weights and biases can be calculated. For the weights:

$$\frac{\partial L}{\partial w_{ij}^{(l)}} = \frac{\partial L}{\partial z_j^{(l)}} \cdot \frac{\partial z_j^{(l)}}{\partial w_{ij}^{(l)}}, \tag{1.8}$$

and for the biases as:

$$\frac{\partial L}{\partial b_j^{(l)}} = \frac{\partial L}{\partial z_j^{(l)}} \cdot \frac{\partial z_j^{(l)}}{\partial b_j^{(l)}}. \tag{1.9}$$

Since $z_j^{(l)} = \sum_{i=1}^{n_{l-1}} w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)}$, we have:

$$\frac{\partial z_j^{(l)}}{\partial w_{ij}^{(l)}} = a_i^{(l-1)} \tag{1.10}$$

$$\frac{\partial z_j^{(l)}}{\partial b_j^{(l)}} = 1 \tag{1.11}$$

So the gradients become:

$$\frac{\partial L}{\partial w_{ij}^{(l)}} = \frac{\partial L}{\partial z_j^{(l)}} \cdot a_i^{(l-1)} \tag{1.12}$$

$$\frac{\partial L}{\partial b_j^{(l)}} = \frac{\partial L}{\partial z_j^{(l)}} \tag{1.13}$$

To compute $\frac{\partial L}{\partial z_j^{(l)}}$, we use the chain rule and work our way back from the output layer:

$$\frac{\partial L}{\partial z_j^{(l)}} = \sum_{k=1}^{n_{l+1}} \frac{\partial L}{\partial z_k^{(l+1)}} \cdot \frac{\partial z_k^{(l+1)}}{\partial a_j^{(l)}} \cdot \frac{\partial a_j^{(l)}}{\partial z_j^{(l)}} \tag{1.14}$$

Here, $\frac{\partial z_k^{(l+1)}}{\partial a_j^{(l)}} = w_{jk}^{(l+1)}$ and for the sigmoid activation function, $\frac{\partial a_j^{(l)}}{\partial z_j^{(l)}} = a_j^{(l)}(1-a_j^{(l)})$. Therefore,

$$\frac{\partial L}{\partial z_j^{(l)}} = \left( \sum_{k=1}^{n_{l+1}} \frac{\partial L}{\partial z_k^{(l+1)}} w_{jk}^{(l+1)} \right) a_j^{(l)}(1 - a_j^{(l)}) \tag{1.15}$$

- **Update weights and biases:** Using the computed gradients, we update the weights and biases with the learning rate $\eta$:

$$w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \eta \frac{\partial L}{\partial w_{ij}^{(l)}}, \tag{1.16}$$

$$b_j^{(l)} \leftarrow b_j^{(l)} - \eta \frac{\partial L}{\partial b_j^{(l)}}. \tag{1.17}$$

This process of forward propagation, loss calculation, backpropagation, and weight updates is repeated for multiple epochs or until the model converges to an optimal solution. The weights and biases of the network are iteratively updated based on the computed gradients of the loss function during backpropagation. In summary, a feedforward neural network learns to predict binary classes by minimizing the loss function using gradient descent optimization. The learning process involves the following steps:

1. Defining the network architecture and initializing the weights and biases.
2. Choosing the activation functions for the different layers, e.g., sigmoid, ReLU, or tanh.
3. Performing forward propagation to compute the predicted outputs.
4. Calculating the loss using a suitable loss function, such as binary cross-entropy.
5. Computing the gradients of the loss function with respect to the weights and biases through backpropagation.
6. Updating the weights and biases using the computed gradients and a learning rate.
7. Repeating steps 3-6 for multiple epochs or until convergence. Convergence in machine learning training refers to the point at which the model's performance on the validation or test set stops improving or plateaus, indicating that the model has learned as much as it can from the available data.

The feedforward neural network adjusts its parameters during training to minimize the difference between the predicted outputs and actual labels, allowing it to make

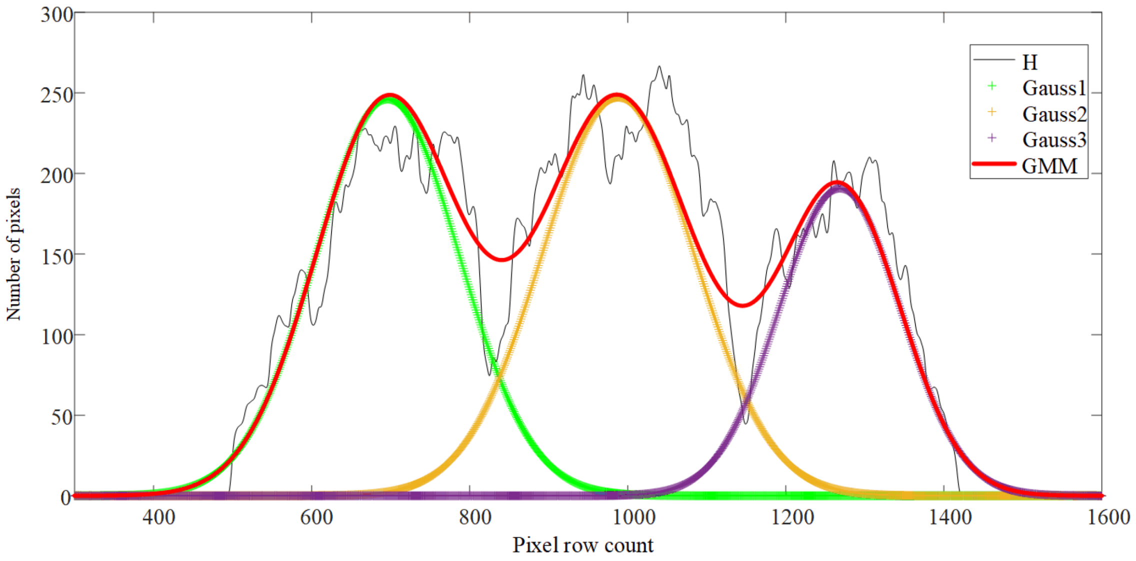


Figure 1.13: Example of a Gaussian Mixture Model (GMM) Distribution. The red curve is the GMM; the black curve is the input data (H); and Gauss 1, Gauss 2, and Gauss 3 (the "+" curve) are the components of the GMM.[63]

accurate binary predictions on unseen data. An example of a one-hidden-layer neural network architecture is shown in Fig. 1.12.

### *1.3.4 Gaussian Mixture Modelling for Probability Density Estimation*

Probability density estimation is a fundamental problem in statistics and data analysis, which involves estimating the probability density function of a random variable from a sample of data. Gaussian mixture modelling is a popular technique for probability density estimation that involves fitting a mixture of Gaussian distributions to the data. This technique is employed in chapter 2.

Let X = x1, x2, ..., xn be a set of n observations of a continuous random variable X. The goal of probability density estimation is to estimate the probability density function p(x) of X from the observed data. Gaussian mixture modeling is a technique that models the probability density function of X as a weighted sum of K Gaussian distributions, where K is the number of components in the mixture model. An example is shown in Fig. 1.13 with 3 estimated components in the GMM model.[63] The probability density function of X can be written as:

$$p(x) = \sum_{k=1}^{K} \pi_k \phi(x|\mu_k, \Sigma_k) \tag{1.18}$$

where $\pi_k$ is the mixing coefficient of the kth Gaussian component such that $0 \geq \pi_k \geq 1$ for all k=1,...K,

$$\sum_{k=1}^{K} \pi_k = 1 \tag{1.19}$$

$\mu_k$ is the mean vector of the k-th Gaussian component, $\Sigma_k$ is the covariance matrix of the k-th Gaussian component, and $\phi(x|\mu_k, \Sigma_k)$ is the probability density function of a Gaussian distribution with mean $\mu_k$ and covariance $\Sigma_k$, evaluated at x.

The goal of Gaussian mixture modeling is to estimate the parameters of the mixture model, which include the mixing coefficient, mean and covariance of each Gaussian component. The parameters can be estimated using the Expectation-Maximization (EM) algorithm, which is an iterative algorithm that alternates between the E-step and the M-step.

In the E-step, the posterior probabilities (or "responsibilities") of each data point $x_i$ belonging to each Gaussian component k are computed, given the current estimates of the parameters:

$$\gamma_k(x_i) = \frac{\pi_k \phi(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \phi(x_i|\mu_j, \Sigma_j)} \tag{1.20}$$

where $\gamma_k^{x_i}$ is the posterior probability of $x_i$ belonging to the k-th Gaussian component.

In the M-step, the parameters of the mixture model are re-estimated using the posterior probabilities computed in the E-step:

$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} \gamma_k^{x_i} \tag{1.21}$$

$$\mu_k = \frac{\sum\limits_{i=1}^{n} \gamma_k^{x_i} x_i}{\sum\limits_{i=1}^{n} \gamma_k^{x_i}} \tag{1.22}$$

$$\Sigma_k = \frac{\sum\limits_{i=1}^{n} \gamma_k^{x_i} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum\limits_{i=1}^{n} \gamma_k^{x_i}} \tag{1.23}$$

where $\pi_k$, $\mu_k$ and $\Sigma_k$ are the updated mixing coefficient, mean and covariance of the k-th Gaussian component, respectively. In the E-Step, the gaussian parameters are kept fixed, where as the assignments are updated. In the M-Step, the assignments are kept fixed, where as the parameters of the distribution are updated. The iterative algorithm converges when the update in parameters of the mixture components is minimal.

In summary, Gaussian mixture modeling is a powerful technique for probability density estimation that models the probability density function of a random variable as a weighted sum of Gaussian distributions. The EM algorithm is used to estimate the parameters of the mixture model, and the resulting model can be used to estimate the probability density function of the data.

# REFERENCES

[1] Leandro C Tabares, Ankur Gupta, Thijs J Aartsma, and Gerard W Canters. Tracking electrons in biological macromolecules: From ensemble to single molecule. *Molecules*, 19(8):11660–11678, 2014.

[2] Boris Rotman. Measurement of activity of single molecules of $\beta$-D-galactosidase. *Proceedings of the National Academy of Sciences*, 47(12):1981–1991, 1961.

[3] Ashok A Deniz, Samrat Mukhopadhyay, and Edward A Lemke. Single-molecule biophysics: at the interface of biology, physics and chemistry. *Journal of the Royal Society Interface*, 5(18):15–45, 2008.

[4] Shelley A Claridge, Jeffrey J Schwartz, and Paul S Weiss. Electrons, photons, and force: quantitative single-molecule measurements from physics to biology. *ACS Nano*, 5(2):693–729, 2011.

[5] Massimiliano Di Ventra and Masateru Taniguchi. Decoding DNA, RNA and peptides with quantum tunnelling. *Nature Nanotechnology*, 11(2):117–126, 2016.

[6] William E Moerner and Michel Orrit. Illuminating single molecules in condensed matter. *Science*, 283(5408):1670–1676, 1999.

[7] Ben NG Giepmans, Stephen R Adams, Mark H Ellisman, and Roger Y Tsien. The fluorescent toolbox for assessing protein location and function. *Science*, 312 (5771):217–224, 2006.

[8] Joshua Hihath, Bingqian Xu, Peiming Zhang, and Nongjian Tao. Study of single-nucleotide polymorphisms by means of electrical conductance measurements. *Proceedings of the National Academy of Sciences*, 102(47):16979–16983, 2005.

[9] Xiaoyin Xiao, Bingqian Xu, and Nongjian Tao. Changes in the conductance of single peptide molecules upon metal-ion binding. *Angewandte Chemie International Edition*, 43(45):6148–6152, 2004.

[10] Fang Chen, Joshua Hihath, Zhifeng Huang, Xiulan Li, and NJ Tao. Measurement of single-molecule conductance. *Annu. Rev. Phys. Chem.*, 58:535–564, 2007.

[11] Amanda M Moore and Paul S Weiss. Functional and spectroscopic measurements with scanning tunneling microscopy. *Annu. Rev. Anal. Chem.*, 1:857–882, 2008.

[12] Gavin David Scott and Douglas Natelson. Kondo resonances in molecular devices. *ACS Nano*, 4(7):3560–3579, 2010.

[13] T Albrecht. Electrochemical tunnelling sensors and their potential applications. *Nature Communications*, 3(1):829, 2012.

[14] Wikimedia Commons. File:rastertunnelmikroskop-schema.svg — wikimedia commons, the free media repository, 2020. URL `https://commons.wikimedia.org/w/index.php?title=File:Rastertunnelmikroskop-schema.svg&oldid=456882138`.

[15] Francesca Moresco. Manipulation of large molecules by low-temperature STM: model systems for molecular electronics. *Physics Reports*, 399(4):175–225, 2004.

[16] Supriyo Datta, Weidong Tian, Seunghun Hong, R Reifenberger, Jason I Henderson, and Clifford P Kubiak. Current-voltage characteristics of self-assembled monolayers by scanning tunneling microscopy. *Physical Review Letters*, 79(13): 2530, 1997.

[17] Shuai Chang, Jin He, Ashley Kibel, Myeong Lee, Otto Sankey, Peiming Zhang,

and Stuart Lindsay. Tunnelling readout of hydrogen-bonding-based recognition. *Nature Nanotechnology*, 4(5):297–301, 2009.

[18] Yanan Zhao, Brian Ashcroft, Peiming Zhang, Hao Liu, Suman Sen, Weisi Song, JongOne Im, Brett Gyarfas, Saikat Manna, Sovan Biswas, et al. Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nature Nanotechnology*, 9(6):466–473, 2014.

[19] Abraham Nitzan. *Chemical dynamics in condensed phases: relaxation, transfer and reactions in condensed molecular systems*. Oxford university press, 2006.

[20] Christopher D Bostick, Sabyasachi Mukhopadhyay, Israel Pecht, Mordechai Sheves, David Cahen, and David Lederman. Protein bioelectronics: a review of what we do and do not know. *Reports on Progress in Physics*, 81(2):026601, 2018.

[21] Ramesh Y Adhikari, Nikhil S Malvankar, Mark T Tuominen, and Derek R Lovley. Conductivity of individual Geobacter pili. *RSC Advances*, 6(10):8354–8357, 2016.

[22] F. J. R. Meysman, R. Cornelissen, S. Trashin, R. Bonne, S. H. Martinez, J. van der Veen, C. J. Blom, C. Karman, J. L. Hou, R. T. Eachambadi, J. S. Geelhoed, K. Wael, H. J. E. Beaumont, B. Cleuren, R. Valcke, H. S. J. van der Zant, H. T. S. Boschker, and J. V. Manca. A Highly Conductive Fibre Network Enables Centimetre-Scale Electron Transport in Multicellular Cable Bacteria. *Nat. Commun.*, 10:4120, 2019.

[23] N. Amdursky, D. Marchak, L. Sepunaru, I. Pecht, M. Sheves, and D. Cahen. Electronic Transport via Proteins. *Adv. Mater.*, 26:7142, 2014.

[24] G. Vattay, D. Salahub, I. a. Csabai, A. Nassimi, and S. A. Kaufmann. Quantum Criticality at the Origin of Life. *J. Phys.: Conf. Ser.*, 626:012023, 2015.

[25] Stuart Lindsay. Ubiquitous Electron Transport in Non-Electron Transfer Proteins. *Life*, 10(5), 2020.

[26] Jesús E Contreras-Naranjo and Oscar Aguilar. Suppressing non-specific binding of proteins onto electrode surfaces in the development of electrochemical immunosensors. *Biosensors*, 9(1):15, 2019.

[27] B. Zhang, W. Song, P. Pang, Y. Zhao, P. Zhang, I. Csabai, G. Vattay, and S. Lindsay. Observation of Giant Conductance Fluctuations in a Protein. *Nano Futures*, 1:035002, 2017.

[28] Bintian Zhang, Weisi Song, Pei Pang, Huafang Lai, Qiang Chen, Peiming Zhang, and Stuart Lindsay. Role of contacts in long-range protein conductance. *Proceedings of the National Academy of Sciences*, 116(13):5886–5891, 2019.

[29] B. Zhang and S. Lindsay. Electronic Decay Length in a Protein Molecule. *Nano Lett.*, 19:4017, 2019.

[30] Bintian Zhang, Weisi Song, Jesse Brown, Robert Nemanich, and Stuart Lindsay. Electronic Conductance Resonance in Non-Redox-Active Proteins. *Journal of the American Chemical Society*, 142(13):6432–6438, 2020.

[31] Adi Salomon, David Cahen, Stuart Lindsay, John Tomfohr, Vincent B Engelkes, and C Daniel Frisbie. Comparison of electronic transport measurements on organic molecules. *Advanced Materials*, 15(22):1881–1890, 2003.

[32] Q. Lu, K. Liu, H. Zhang, Z. Du, X. Wang, and F. Wang. From Tunneling to Hopping: A Comprehensive Investigation of Charge Transport Mechanism in

Molecular Junctions Based on Oligo(p-phenylene ethynylene)s. *ACS Nano*, 3: 3861, 2009.

[33] Bintian Zhang, Eathen Ryan, Xu Wang, Weisi Song, and Stuart Lindsay. Electronic Transport in Molecular Wires of Precisely Controlled Length Built from Modular Proteins. *ACS Nano*, 16(1):1671–1680, 2022.

[34] Catharine Shipps, H. Ray Kelly, Peter J. Dahl, Sophia M. Yi, Dennis Vu, David Boyer, Calina Glynn, Michael R. Sawaya, David Eisenberg, Victor S. Batista, and Nikhil S. Malvankar. Intrinsic electronic conductivity of individual atomically resolved amyloid crystals reveals micrometer-long hole hopping via tyrosines. *Proceedings of the National Academy of Sciences*, 118(2):e2014139118, 2021.

[35] L. Sepunaru, I. Pecht, M. Sheves, and D. Cahen. Solid-State Electron Transport across Azurin: From a Temperature-Independent to a Temperature-Activated Mechanism. *J. Am. Chem. Soc.*, 133:2421, 2011.

[36] K. S. Kumar, R. R. Pasula, S. Lim, and C. A. Nijhuis. Long-Range Tunneling Processes across Ferritin-Based Junctions. *Adv. Mater.*, 28:1824, 2016.

[37] B. Kayser, J. A. Fereiro, R. Bhattacharyya, S. R. Cohen, A. Vilan, I. Pecht, M. Sheves, and D. Cahen. Solid-State Electron Transport via the Protein Azurin is Temperature-Independent Down to 4 K. *journal of Physical Chemistry Letters*, 11:144, 2020.

[38] S. Mukhopadhyay, S. Dutta, I. Pecht, M. Sheves, and D. Cahen. Conjugated Cofactor Enables Efficient Temperature-Independent Electronic Transport Across approximately 6 nm Long Halorhodopsin. *J. Am. Chem. Soc.*, 137:11226, 2015.

[39] K. Garg, M. Ghosh, T. Eliash, J. H. van Wonderen, J. N. Butt, L. Shi, X. Jiang,

F. Zdenek, J. Blumberger, I. Pecht, M. Sheves, and D. Cahen. Direct evidence for heme-assisted solid-state electronic conduction in multi-heme c-type cytochromes. *Chem. Sci.*, 9:7304, 2018.

[40] K. Garg, S. Raichlin, T. Bendikov, I. Pecht, M. Sheves, and D. Cahen. Interface Electrostatics Dictates the Electron Transport via Bioelectronic Junctions. *ACS Appl. Mater. Interfaces*, 10:41599, 2018.

[41] J. A. Fereiro, X. Yu, I. Pecht, M. Sheves, J. C. Cuevas, and D. Cahen. Tunneling explains efficient electron transport via protein junctions. *Proc. Natl. Acad. Sci. U. S. A.*, 115:E4577, 2018.

[42] O. E. Castañeda Ocampo, P. Gordiichuk, S. Catarci, D. A. Gautier, A. Herrmann, and R. C. Chiechi. Mechanism of Orientation-Dependent Asymmetric Charge Transport in Tunneling Junctions Comprising Photosystem I. *J. Am. Chem. Soc.*, 137:8419, 2015.

[43] Q. Chi, O. Farver, and J. Ulstrup. Long-range protein electron transfer observed at the single-molecule level: In situ mapping of redox-gated tunneling resonance. *Proc. Natl. Acad. Sci. U. S. A.*, 102:16203, 2005.

[44] A. Alessandrini, S. Corni, and P. Facci. Unravelling single metalloprotein electron transfer by scanning probe techniques. *Phys. Chem. Chem. Phys.*, 8:4383, 2006.

[45] E. A. Pia, Q. Chi, D. D. Jones, J. E. Macdonald, J. Ulstrup, and M. Elliott. Single-molecule mapping of long-range electron transport for a cytochrome b(562) variant. *Nano Lett.*, 11:176, 2011.

[46] N. J. Tao. Probing Potential-Tuned Resonant Tunneling through Redox Molecules with Scanning Tunneling Microscopy. *Phys. Rev. Lett.*, 76:4066, 1996.

[47] J. M. Artes, I. Diez-Perez, and P. Gorostiza. Transistor-like behavior of single metalloprotein junctions. *Nano Lett.*, 12:2679, 2012.

[48] A. M. Kuznetsov and J. Ulstrup. Single-molecule electron tunnelling through multiple redox levels with environmental relaxation. *J. Electroanal. Chem.*, 564: 209, 2004.

[49] I. V. Pobelov, Z. Li, and T. Wandlowski. Electrolyte gating in redox-active tunneling junctions–an electrochemical STM approach. *J. Am. Chem. Soc.*, 130: 16045, 2008.

[50] Sara H Mejías, Begoña Sot, Raul Guantes, and Aitziber L Cortajarena. Controlled nanometric fibers of self-assembled designed protein scaffolds. *Nanoscale*, 6(19):10982–10988, 2014.

[51] J. Jortner, M. Bixon, T. Langenbacher, and M. E. Michel-Beyerle. Charge transfer and transport in DNA. *Proc. Natl. Acad. Sci. U. S. A.*, 95:12759, 1998.

[52] Sara H Mejias, Antonio Aires, Pierre Couleaud, and Aitziber L Cortajarena. Designed repeat proteins as building blocks for nanofabrication. *Protein-based Engineered Nanostructures*, pages 61–81, 2016.

[53] Ewan RG Main, Yong Xiong, Melanie J Cocco, Luca D'Andrea, and Lynne Regan. Design of stable $\alpha$-helical arrays from an idealized TPR motif. *Structure*, 11(5):497–508, 2003.

[54] Tommi Kajander, Aitziber L Cortajarena, Simon Mochrie, and Lynne Regan. Structure and stability of designed TPR protein superhelices: unusual crystal packing and implications for natural TPR proteins. *Acta Crystallographica Section D: Biological Crystallography*, 63(7):800–811, 2007.

[55] Jie Liu, Qi Zheng, Yiqun Deng, Neville R Kallenbach, and Min Lu. Conformational transition between four and five-stranded phenylalanine zippers determined by a local packing interaction. *Journal of Molecular Biology*, 361(1): 168–179, 2006.

[56] Jie Liu, Wei Yong, Yiqun Deng, Neville R Kallenbach, and Min Lu. Atomic structure of a tryptophan-zipper pentamer. *Proceedings of the National Academy of Sciences*, 101(46):16156–16161, 2004.

[57] S. Kamtekar, A. J. Berman, J. Wang, J. M. Lazaro, M. de Vega, L. Blanco, M. Salas, and T. A. Steitz. Insights into Strand Displacement and Processivity from the Crystal Structure of the Protein-Primed DNA Polymerase of Bacteriophage Phi29. *Mol. Cell*, 16:609, 2004.

[58] L. Blanco, A. Bernad, J. M. Lazaro, G. Martin, C. Garmendia, and M. Salas. Highly Efficient DNA Synthesis by the Phage Phi 29 DNA Polymerase: Symmetrical Mode of DNA Replication. *J. Biol. Chem.*, 264:8935, 1989.

[59] J. A. Esteban, M. Salas, and L. Blanco. Fidelity of Phi 29 DNA Polymerase. Comparison Between Protein-Primed Initiation and DNA Polymerization. *J. Biol. Chem.*, 268:2719, 1993.

[60] Medhanie Tesfay Gebrekidan, Christian Knipfer, Florian Stelzle, Juergen Popp, Stefan Will, and Andreas Braeuer. A shifted-excitation Raman difference spectroscopy (SERDS) evaluation strategy for the efficient isolation of Raman spectra from extreme fluorescence interference. *Journal of Raman Spectroscopy*, 47(2): 198–209, 2016.

[61] Yan Zhu, Shaghayegh Gharghabi, Diego Furtado Silva, Hoang Anh Dau, Chin-Chia Michael Yeh, Nader Shakibay Senobari, Abdulaziz Almaslukh, Kaveh

Kamgar, Zachary Zimmerman, Gareth Funning, Abdullah Mueen, and Eamonn Keogh. The Swiss army knife of time series data mining: ten useful things you can do with the matrix profile and ten lines of code. *Data Mining and Knowledge Discovery*, 34:949–979, 2020.

[62] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[63] Zhenzhen Cheng, Lijun Qi, Yifan Cheng, Yalei Wu, and Hao Zhang. Interlacing Orchard Canopy Separation and Assessment using UAV Images. *Remote Sensing*, 12(5), 2020.

Chapter 2

# EXPLORING SOLVENT EFFECTS ON ELECTRONIC TRANSPORT IN PROTEINS: A STUDY OF A SERIES OF CONSENSUS TETRATRICOPEPTIDE REPEAT (CTPR) PROTEINS AND ZIPPER PROTEINS

*In this chapter, the protein design and preparations were done by Dr. Eathen Ryan and Dr. Nicholas Halloran. The STM measurements were carried out by Dr. Sepideh Afsari. My contribution included the algorithm design and the statistical analysis of the experimental data.*

## 2.1 Introduction

Proteins possess exceptional electronic properties, [1–5] making them a promising material for creating self-assembling molecular electronic components at the amino - acid residue scale.[6] External electric fields can influence their performance as nanoscale conductors in a polarizable medium, and their charge transport is affected by solvent dynamics. [7] One way to regulate solvent dynamics without compromising protein function is by using heavy water as a solvent. Heavy water is denser, more viscous, and has higher boiling and melting temperatures than regular water. However, it can be lethal at high concentrations.[8] At the single-molecule level, $D_2O$ stiffens proteins due to the damping of fluctuations and not because of strong intramolecular D bonding.[9]

$D_2O$ significantly impacts enzymatic processes that involve proton tunneling [10] and alters electron transfer reaction rates.[11–14] The mechanism of this isotope effect is debatable. Changes in protein flexibility [9] affect the spectral distribution of fluctuations and alter the activation barrier for electron transfer.[12] On the other hand, recent calculations for azurin in $D_2O$ indicate that the barrier does not change and suggest that the exponential pre-factor (i.e., the attempt frequency) is most affected.[15] Recent studies have used a scanning probe microscope to measure electron transport in single protein molecules,[5, 16–19] with notable results in oligomeric chains of cytochrome proteins OmCs that showed efficient electron transport over micron distances.[20] A recent study found a significant isotope effect in OmcS nanowires, with a  200-fold decrease in conductivity observed on introducing $D_2O$ at room temperature.[21]

This study investigates the effects of $D_2O$ on the electronic conductance of a series of proteins without redox cofactors. The measurement is performed as a function of the length to distinguish between isotope-induced changes in bulk diffusion constant and those affecting contact resistance. The researchers also synthesized and measured zipper proteins consisting of a five-helix bundle with either tryptophan or phenylalanine residues. Tryptophan contains indole rings with a labile proton that readily oxidizes, [22] while phenylalanine contains phenyl rings that lack hydrogen/deuterium binding sites and is not readily oxidized. Studies of the single molecule conductance of these proteins are described in this chapter.

## 2.2 Experimental Design

### *2.2.1 Preparation of STM Substrates and STM Probes*

**Probe Etching:** Following the method described earlier for gold probes [5, 23], a 0.25 mm diameter Palladium wire was lowered into an etching solution of 1:1 HCl and EtOH using a precision mechanical translation stage. A 5 kHz square wave with a peak-to-peak voltage of 40 V was applied between the Palladium wire and a platinum foil counter electrode. The wire was then lowered into the etching solution until the RMS current reached 400 mA, which corresponded to an immersion depth of 2 mm-3 mm. Etching was continued at constant peak-to-peak voltage until the current fell to zero.

In the second step, the tip was withdrawn from the solution, and the voltage was reduced to approximately 12 V peak to peak. The Pd wire was then quickly immersed in the same solution until it reached a peak current of 40-80 mA, taking approximately 0.25 seconds, and immediately pulled out of the etching solution. This two-step process produced a probe shape with a sharp apex on top of a cone with a half-angle of approximately 25° ( A gold STM probe is shown in the optical micrograph in Fig. 2.1(a) as an example).

The broad cone was required to support the high-density polyethylene (HDPE) coating while the sharp apex penetrated it. It was crucial that the probe surface was smooth and the apex was sharp without any visible end resolvable under 250× magnification. An example of a sharp probe (Gold) that could not be adequately insulated is shown in Fig. 2.1(b). Etching was repeated until the desired shape was achieved, either from the first (40 V) or the second (12 V) step, depending on whether the overall shape was bad (Fig. 2.1(b)) or the end just needed further sharpening. Finally, satisfactory probes were rinsed with ethanol (EtOH) and dried using nitrogen

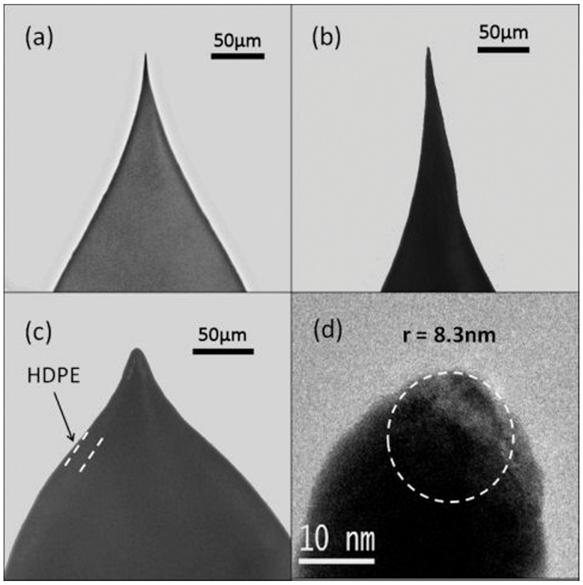


Figure 2.1: Profiles of the etched and insulated gold probes [23]. Optical images under 250× magnification of (a) a good etched STM probe (b) a poor etched gold STM probe, and (c) a good coated STM probe. A TEM image (d) of a typical good STM probe with radius of curvature equal to 8.3 nm in this case.

($N_2$) gas.

**Probe Coating:** Following the method described earlier for gold probes [5, 23], etched Pd probes were immersed in a 1:3 $H_2O_2$ and $H_2SO_4$ solution (the "piranha" solution) for 1-2 minutes. The probes were then rinsed with $H_2O$ and EtOH and air-dried with $N_2$ gas. High-density polyethylene (HDPE) granules from Alfa Aesar, which contained no traceable amounts of plasticizers, were melted onto the platform of the tip coating instrument. The instrument was previously heated to 270-280 °C. Once the polymer melt became transparent, the Pd tip was pushed up through the molten blob at a rate of 30 $\mu$m/s - 100 $\mu$m/s. As the HDPE began to harden on the apex of the probe, which was now above the reservoir of melted polymer, the Pd tip was withdrawn to below the level of the melted HDPE, and the first step was repeated.

A high-quality scanning tunneling microscopy (STM) probe had an HDPE coating that came to a point at the apex with no visible protrusions. The coating should be smooth and continuous along the length of the probe, and the Pd probe within the HDPE must be straight to the top of the coating. The coating should be at least 5 $\mu$m thick near the apex of the tip and thicker about the remainder of the probe, an example for the gold probe is shown in Fig. 2.1(c).

A low-quality tip may have visible sharp protrusions from the apex of the coating. If the tip was still sharp, re-coating helped to reduce the amount of visible protrusion. A poor quality tip may also be bent, such that the exposure of the tip was not at the top of the coating, or it was leaving it completely insulated. Such probes were discarded. Each probe was tested by STM in 1 mM PB buffer at -0.5-V bias to ensure the leakage current was <1 pA.[5]

**Substrate Preparation and Functionalization:** Maintaining the established protocol, [19] initially, 200 nm of Pd was deposited on a 10 nm Cr adhesion layer onto four-inch p-type Si wafers, followed by the deposition using an e-beam evaporator (Lesker PVD 75). To ensure effective functionalization and prevent aggregation of CTPR proteins, a solution of 50 mM tris buffer pH 7.4 containing 10 mM TCEP, 200 mM $NaCl$, 10 mM $(NH_4)_2SO_4$, and $MgCl_2$ was used. Monolayers of CTPR4, CTPR8, CTPR12, and CTPR16 series were prepared by immersing the chip in 20, 50, 30, and 100 $\mu$M solutions, respectively, for approximately 16 hours.

After functionalization, the chip was removed, rinsed with water, blown dry with nitrogen, and immediately used. The monolayer's thickness was measured on a Gaertner L 123b Ellipsometer (Gaerner Scientific Corporation) using a refractive index of 1.5 for a thin organic layer. Five measurements were conducted on different chip locations to calculate the monolayer thickness's mean value.

The zipper proteins were only available in 1 $\mu$M concentrations, and the film thicknesses measured by ellipsometry were 4.2 ± 0.2 nm (Trp-Zipper), 4.1 ± 0.4 nm

(Phe-Zipper) and 3.7 ± 0.3 nm (Phe V2 Zipper). These values are greater than the width of the proteins ( 2 nm), so the molecules are likely tilted, similar to the results obtained with CTPR proteins at this lower concentration.[19]

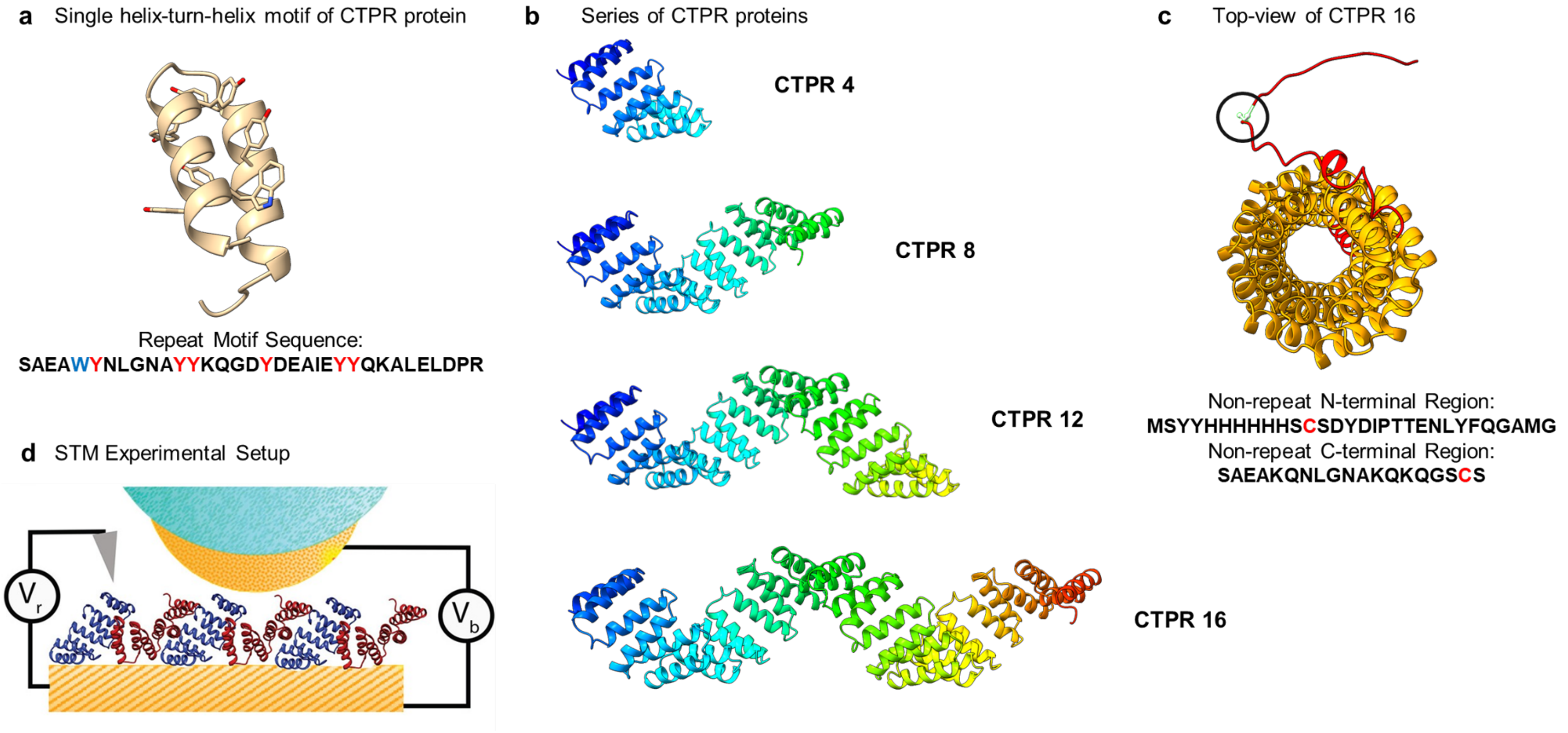

a Single helix-turn-helix motif of CTPR protein
Repeat Motif Sequence:
SAEAWYNLGNAYYKQGDYDEAIEYYQKALELDPR
b Series of CTPR proteins
CTPR 4
CTPR 8
CTPR 12
CTPR 16
c Top-view of CTPR 16
Non-repeat N-terminal Region:
MSYYHHHHHHSCSDYDIPTTENLYFQGAMG
Non-repeat C-terminal Region:
SAEAKQNLGNAKQKQGSCS
d STM Experimental Setup
Vr
Vb

Figure 2.2: Consensus Tetratricopeptide (CTPR) Wires and the Conductance Measurement Setup. (a) Aromatic residues in a single helix-turn-helix motif of the CTPR protein. The repeat motif sequence of 34 amino acids is shown with Tryptophan (W) and Tyrosine (Y) residues highlighted in blue and red respectively. (b) The core structure of CTPR constructs perpendicular to the long axis showing each repeating motif with a specific color. CTPR wires were synthesized with N- and C- terminal cysteines. (c) A top-view of the CTPR16 shows the helical structure of CTPR proteins. The N-terminal cysteine is marked by a black circle. The sequences of the terminal regions of the CTPR proteins which are not a part of the repeating motif are shown with the cysteine residue highlighted in red. (d) Conductance measurements: Current measured by an STM probe jumps as contact is made to the molecular layer, shown here for CTPR4 molecules docked in an alternating orientation for which the predicted tilt angle is 46°. $V_b$ is the bias applied to the tip-substrate junction, and $V_r$ is the bias applied to a Ag/AgCl reference electrode in the electrolyte solution. Probes are insulated (blue) to within a few nm of their apex, reducing leakage currents to <1pA. The initial set point ($Z_0$) was 4 pA with $V_b$ = 200 mV.[19]

### *2.2.2 Engineering CTPR and Zipper Proteins*

#### *2.2.2.1 Consensus Tetratricopeptide Repeat (CTPR) protein expression*

As described in Section 1.2.1 CTPR proteins represent a standardized 34 amino acid (Fig. 2.2) helix-turn-helix repeat module that can be combined in tandem to create proteins with various repeat numbers, from 2 to 20 (CTPR2 to CTPR20). A series of linear proteins were synthesized by concatenation of several helix-turn-helix motifs. The CTPR expression plasmids were derived from wild type tetra-repeat CTPR cloned in the expression vector pPROEX-HTa, and larger constructs were created via digestion and ligation of the core repeat sequence of the CTPR open reading frame using complementary restriction sites BamHI and BglII. This allowed expression of series of CTPRs with increasing numbers of repeats, specifically CTPR 4, 8, 12 and 16. Single cysteine mutations at the N- and C- termini to ensure the formation of chemical attachments to the Palladium (Pd) electrodes (Fig. 2.2). The well-established crystal structures for CTPR8 and CTPR20 show consistency in the arrangement of atoms in the repeat unit [24]. Hence, an accurate extrapolation of the structures for CTPR 4, 8, 12, 16, and 20 can be derived where each repeat corresponds to a length of 0.94 ± 0.07 nm. Mejias et al.[25] demonstrated that the hydrodynamic radius of CTPR polymers increased with the number of repeats, and ellipsometric measurements of CTPR monolayers formed on palladium substrates showed that the film thickness increased linearly with the number of repeat units, ruling out denaturation of proteins on the electrode or flat-lying molecules on the substrate. By plotting the height of CTPR monolayers against their crystallographic length, with the resulting slope of 0.67 ± 0.01, Zhang et al. [19] deduced that the molecules are tilted in the surface monolayer, similar to the tilt reported for CTPR20 monolayer [26]. Using the HADDOCK docking program [27], it was determined that close packing

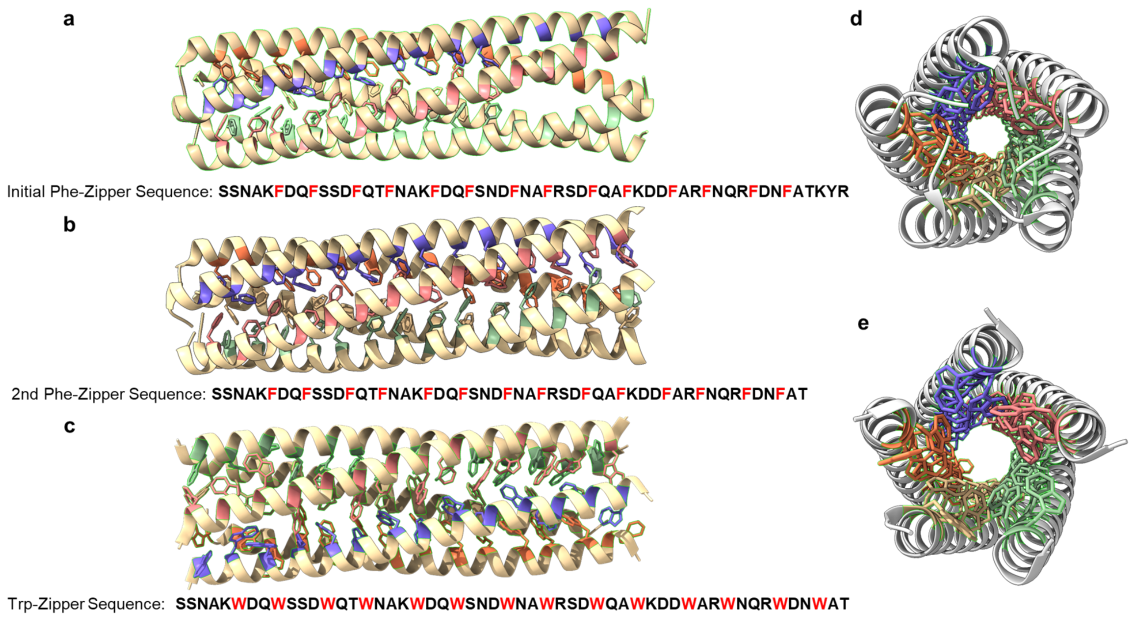

a
Initial Phe-Zipper Sequence: SSNAKFDQFSSDFQTFNAKFDQFSNDFNAFRSDFQAFKDDFARFNQRFDNFATKYR
b
2nd Phe-Zipper Sequence: SSNAKFDQFSSDFQTFNAKFDQFSNDFNAFRSDFQAFKDDFARFNQRFDNFAT
c
Trp-Zipper Sequence: SSNAKWDQWSSDWQTWNAKWDQWSNDWNAWRSDWQAWKDDWARWNQRWDNWAT
d
e

Figure 2.3: Phenylalanine and Tryptophan Zipper Protein Structures. a) Phenylalanine zipper showing the published PDB structure. The sequence highlights the 14 Phenylalanine (F) residues in a single chain. b) A second version (Phe V2) was synthesized with a 3aa deletion to more closely match the length of the Trp-zipper. The sequence highlights the 14 Phenylalanine (F) residues in a single chain. Note the deletion of Lysine (K), Tyrosine (Y), and Arginine (R) residues. c) Tryptophan (Trp) zipper composed of a five helix bundle with N- and C termini modified with cysteines. The sequence highlights the 14 Tryptophan (W) residues in a single chain. d) and e) represent the top-view images of Phe Zipper and Trp Zipper showing the stacking of phenylalanine and tryptophan residues respectively.

is possible with an intermolecular distance of 3.6 nm and an axial displacement of 2.5 nm when neighboring molecules alternate between N-terminal and C-terminal attachments. This model predicted an angle of 46° with respect to the substrate, which is close to the 42° derived from ellipsometric data. The molecules were found to be more upright for monolayers formed from highly concentrated solutions of CTPR as will be seen in Section 2.3.1.

#### *2.2.2.2 Phenylalanine and Tryptophan Zipper protein expression*

The zipper proteins consist of five-helix bundles in which the aromatic residues are stacked in the center to hold the bundle together (Fig. 2.3 (c), tryptophan zipper. PDB: 1T8Z, Fig. 2.3 (a), phenylalanine zipper, PDB: 2GUV). Sequences are listed in Fig. 2.3 (b). For the zipper proteins, synthetic gene fragments encoding for a Phenylalanine - Zipper (PDB 2GUV) and a Tryptophan - Zipper (PDB 1T8Z), including an N-terminal $His_6$, tag followed by a TEV cleavage site, (Genewiz, Inc.), were inserted in Pet27b vectors. A variant of Phe-Zipper containing a 3-residues deletion at the C-terminal was obtained from the original Phe-Zipper gene using complementary primers (Genewiz, Inc.) to match the length of the tryptophan zipper more closely. The plasmids were transformed into *E. coli* BL21(DE3) cells and expressed at 37 °C (induction with IPTG at OD 0.6 to 0.8). The proteins were purified from the soluble fraction by reverse IMAC using a HisTrap 5mL column equilibrated with binding buffer (20 mM sodium phosphate, 40 mM imidazole, and 1 M NaCl at a pH of 7.5); proteins were eluted with 20 mM sodium phosphate, 500 mM imidazole, and 1 M NaCl at pH 7.8. The eluted fractions were desalted on a PD-10 column into 50 mM sodium phosphate and 200 mM NaCl at pH 7.4. The His-tag was removed using TEV protease followed by IMAC. The proteins were treated with 5 mM TCEP and further purified by RP-HPLC on a semi-preparative C18 column. Masses were verified by MALDI. The oligomerization state of the proteins was verified by size exclusion chromatography using a Superdex75 10/300 size exclusion column. More than 95% of each protein eluted at retention times consistent with pentamers. The protein secondary structure was verified by circular dichroism (CD) on a Jasco J-815 CD spectrophotometer. Protein concentration was quantified using UV-Vis A280, and the raw CD spectra were converted to molar ellipticity.

### *2.2.3 STM Conductance Measurements*

As established in prior works, [5, 19] conductances were measured for single molecules (Fig. 2.2 (d)). For data acquisition, STM measurements were performed using a PicoSPM scanning probe microscope (Agilent Technologies) with a DAQ card (PCI-6821 or PCIE-7842R, National Instruments). The buffer solution (1 mM phosphate buffer, pH 7.4) and analytes (CTPR or Zipper protein) were added to the pre-cleaned and sonicated Teflon cell. The substrate was held at 0 V with respect to an Ag/AgCl reference (utilizing a 10 mM salt bridge), and the current between the probe and substrate was monitored as a function of time. The probe was left to stabilize for 2 hours at a setpoint current of 4 pA with a bias of 200 mV before measurement. The servo system was turned off for STM IV sweep measurements, and the probe was retracted by $\Delta Z$ nm (depending on the protein in the gap – see Table 2.2 and Table 2.3) with a speed of 1 nm/s so that the current falls to zero. A jump in the current ($> 40$ pA) was taken to be a signal due to the capture of a molecule, at which point the voltage bias was swept ±200 mV (0.8 s per sweep), and the current was recorded as a function of bias. A molecular conductance was calculated from the slope of each current-vs-voltage (IV) curve. The resulting distributions of conductances were well-fitted by two log-normal distributions. However, as shown in section 2.2.4, artifactual contributions from instrumental drift can dominate the lower conductance feature.

Approximately 500 molecular trapping events were captured in each run, with the gap distance set to $Z_0 + 1$ nm (CTPR4), $Z_0 + 3$ nm (CTPR8), $Z_0 + 4$ nm (CTPR12) and $Z_0 + 5$ nm (CTPR16). $Z_0$ is the gap at a setpoint current of 4 pA with a bias of 200 mV (approximately 2.5 nm [28], but this probably varies significantly depending on the thickness and composition of the monolayer on the surface). Each run was repeated three times, and the distributions were analyzed as described below. As shown elsewhere,[19], the conductance distributions do not change as the gap size was changed within a range that permits molecular trapping. For the zipper proteins, gaps of $Z_0$, $Z_0$+1 nm, and $Z_0$+2 nm were used, finding again that the distribution did not change with gap size. The conductance values and gap sizes are summarized in Table 2.2 and Table 2.3. The peak conductance for each distribution was obtained from the unbinned and auto-filtered data using a Gaussian mixture model (Origin Pro) as explained below.

### *2.2.4 Automatic Filtering of I-V Sweep Data*

Current vs. time data were acquired with the STM servo switched off, so the tip can drift randomly, either away from the surface (which will not trigger the 40 pA threshold) or towards the surface (which will). Our earlier manual filtering was based on selecting I - V curves that were reproducible on sweep - up to sweep - down, but this does not remove artifacts owing to slower drift that leads to a gradual conductance increase over time. Drift can generate a spurious peak at low conductance values. As the probe drifted towards the surface, the current increased, triggering the threshold for recording current (40 pA) without a molecule being contacted. This drift gave rise to a peak-like feature at low conductance, despite the manual selection of data to remove curves affected by drift. The limitation was overcome by developing a fully-automated curve selection procedure. Drift towards the surface gives rise to currents that increase exponentially with time, whereas a stable molecular junction gives a current that is stable with time (Fig. 2.4). Our algorithm identifies and rejects data for which the current increase with time is fitted by a pure exponential. This selection procedure does not affect the higher conductance peak but substantially reduces the number of counts in the lower conductance peak. The remaining events in the lower peak can be attributed to one specific contact with the molecule and one non-specific contact. This chapter presents only auto-filtered data, but a comparison of the distributions obtained with and without auto-filtering is given in Table 2.2 and Table 2.3 and Fig. 2.6). To differentiate statistically between current vs. time signals due to drift and those due to molecular junctions in scanning tunneling microscopy (STM), a 2 - step filtering process was designed and tested based on removing traces in which the peak current increases exponentially with time. This part of the chapter illustrates the algorithm using the experimental I - V traces obtained from the CTPR4

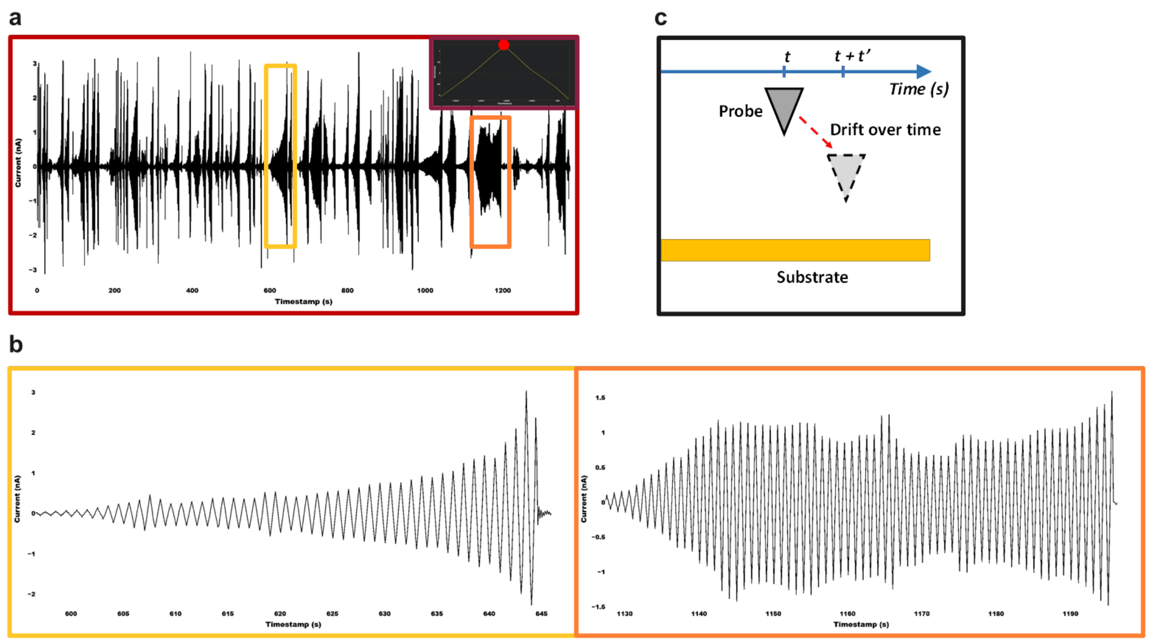


Figure 2.4: Existence of Drift in the STM Conductance Measurements. a) Current changes over time as the probe is stabilized and IV sweeps are recorded from -0.2V to +0.2V with the scale from -3 nA to +3 nA. Each maximum current is selected from IV sweeps lasting 0.8s, with about 1000 IV sweeps in each data gathering run (red circle in Inset). If the gap was stable, this maximum current would be constant. Panel shows b) Exponential and non-exponential (constant amplitude) trend in current can be related to drift (left) and molecular capture (right) respectively. c) Drift is not mono-directional, so the probe can drift towards or across the rough surface, resulting in longer dwell times at a low current.

protein system in the presence of $H_2O$ (Fig. 2.6). The entire workflow is presented in Fig. 2.5.

During a typical experiment, about 1000 I-V scans are triggered. In order to follow the drift, the maximum current during each I-V sweep is recorded, and plots of this maximum current versus time (Fig. 2.5 (a)) clearly show regions of exponential growth in peak current.

The Python script - based automated analysis proceeded as follows: Each recording lasts up to 120 s before the tip is withdrawn and the setpoint reestablished. Data is recorded once the threshold current (40 pA) is exceeded in the withdrawn position.

This typically occurs after 10 to 20 s, so a typical data gathering run lasts for 90 to 100s. During the data gathering, the bias voltage is swept between ± 200mV to gather IV curves. Each sweep (up and down) lasts 0.8s, during which 80000 current data points are recorded. The software locates the peak current in each sweep (Fig. 2.4 (a) and Fig. 2.5 (a) ) and then separates the continuous stream of data into windows (Fig. 2.5 (b)), plots a semi-log plot for the windows (Fig. 2.5 (c)), each one of which corresponds to a data gathering run, for which 80 to 100 peak current events may be recorded. Segments that have an exponentially rising current appear with linear slopes in log(current) plots. This is usually most visible at the end of a data - gathering run, so a linear fit to the last 5 points in a run is made, and $R^2$ is calculated for all the points in a run. Runs for which $R^2 \geq 0.99$ are, therefore, essentially all exponential and are labeled drift ((Fig. 2.5 (d))). Runs that remain ($R^2 < 0.99$) are further kept for further filtration (Fig. 2.5 (e)). This step leaves several runs in which there is clearly drift at the beginning of the run, but the current subsequently becomes stabilized by the formation of a molecular junction. These events were identified by linear fits to the first (lowest in value) 5 points in the plot of log current vs. time (Fig. 2.5 (f)). The mean absolute error (MAE) was then calculated for all of the points in the run. Only those points for which the absolute error exceeded the MAE were kept. Fig. 2.5 (f) shows the linear fit of the least 5 log current points for a sample selected window, and Fig. 2.5 (g) shows the absolute error for all points in the run with the MAE shown as the dashed line. Points above the threshold (MAE = 0.85 in this case) are not fitted by the initial-drift exponential and are kept as molecular junction data for further analysis. The remaining molecular junction data were then fitted using the Gaussian Mixture Model as described in Section 1.3.4. This Bayesian approach optimizes the agreement between modeled data and the actual data set, eliminating the need for binning data.

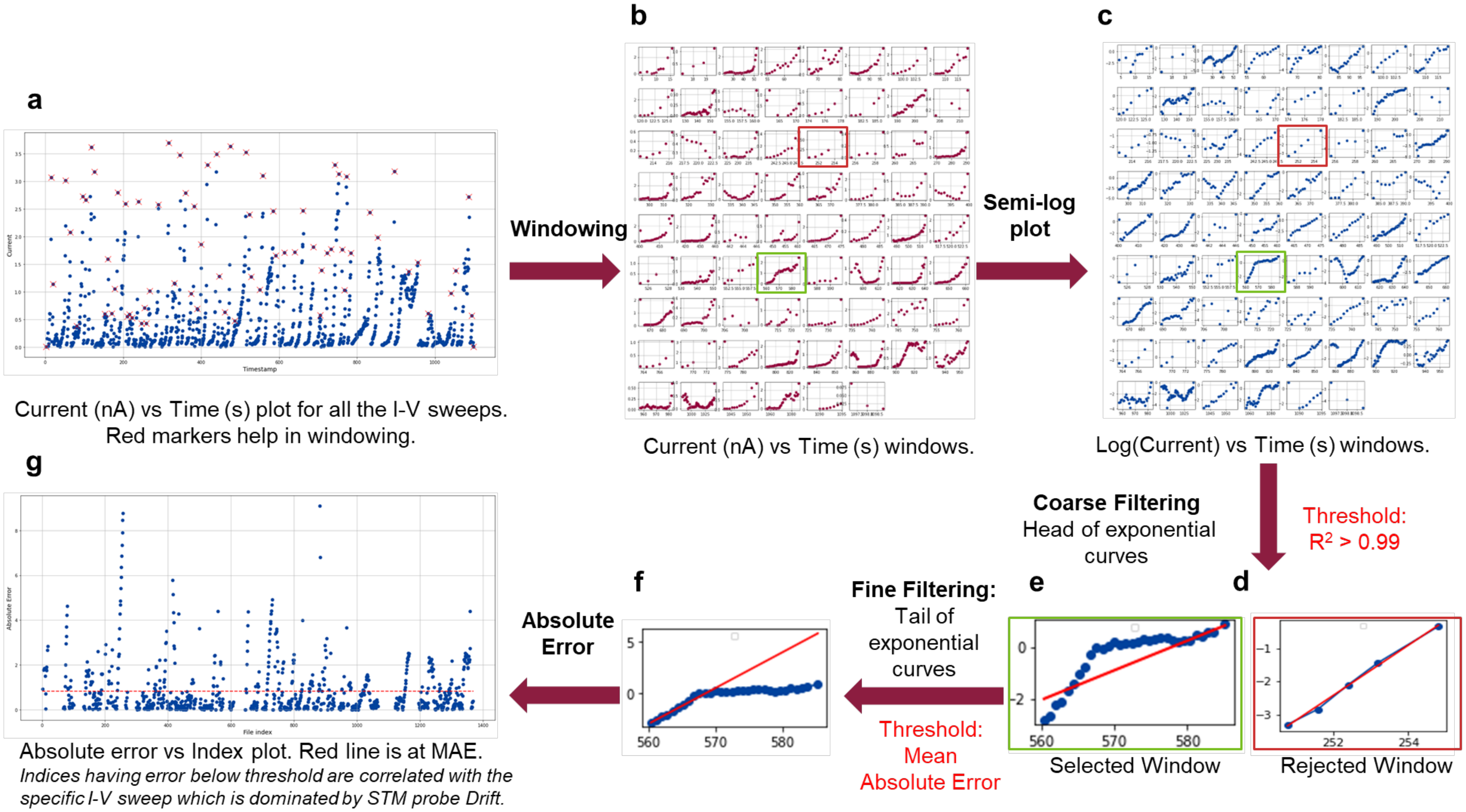

a
Current
Timestamp
Current (nA) vs Time (s) plot for all the I-V sweeps.
Red markers help in windowing.
Windowing
b
Current (nA) vs Time (s) windows.
Semi-log plot
c
Log(Current) vs Time (s) windows.
Coarse Filtering
Head of exponential curves
Threshold:
$R^2 > 0.99$
d
Rejected Window
e
Selected Window
Fine Filtering:
Tail of exponential curves
Threshold:
Mean Absolute Error
f
Absolute Error
g
Absolute Error
File index
Absolute error vs Index plot. Red line is at MAE.
Indices having error below threshold are correlated with the specific I-V sweep which is dominated by STM probe Drift.

Figure 2.5: Workflow of the Custom 2-Step Automated Filtering Algorithm. a) Plot of the maximum current in an IV sweep (scale is 0 to 3 nA) vs timestamp. The red markers show regions of exponential growth and help in the windowing process. Entire current-time trace is broken into windows shown in b) linear current and c) log current scale. Coarse Filtering employs a linear fit to the largest five points of log (current),rejecting those runs for which the $R^2$ exceeded 0.99 for the fit to all the remaining data. An example of a rejected and a selected window is shown in d) and e) respectively (red and green bordered windows are also highlighted in the list of windows shown in b and c). f) Fine filtering computes linear fits to the least five log (current) points. g) Absolute error for each data point in the runs (data concatenated for this plot) with dotted line showing the MAE. Points above this threshold were taken to be molecular junction data.

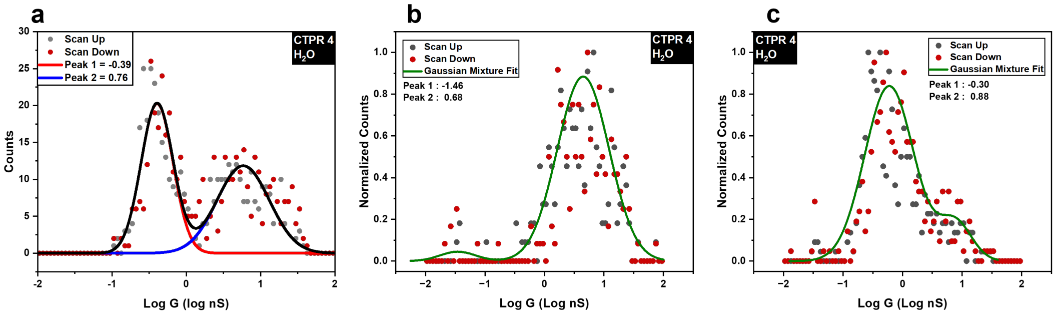


Figure 2.6: Data for CTPR4 in $H_2O$ a) Manually filtered. b) Molecular junction data as selected by the algorithm described above. The green line is the estimated probability density function using the GMM. c) The eliminated "drift" data with GMM fit.

An example of the difference between the manually filtered (reproducible up/down sweeps) and the machine filtering described above is shown in Fig. 2.6 (a-c).

The position of the high conductance peak is little affected in general (Table 2.2 and Table 2.3), but the lower peak is largely eliminated by automated filtering. Control experiments in buffer alone generated spurious data in this low current region at about half the rate observed when molecules are present. Thus, while a significant fraction of these events is entirely spurious, most of the remainder is generated by non-specific contacts for which the current changes with time as the contact point drifts. The slight residue of low-conductance data surviving the filtering process appears as the low-conductance feature in Fig. 2.6 (b).

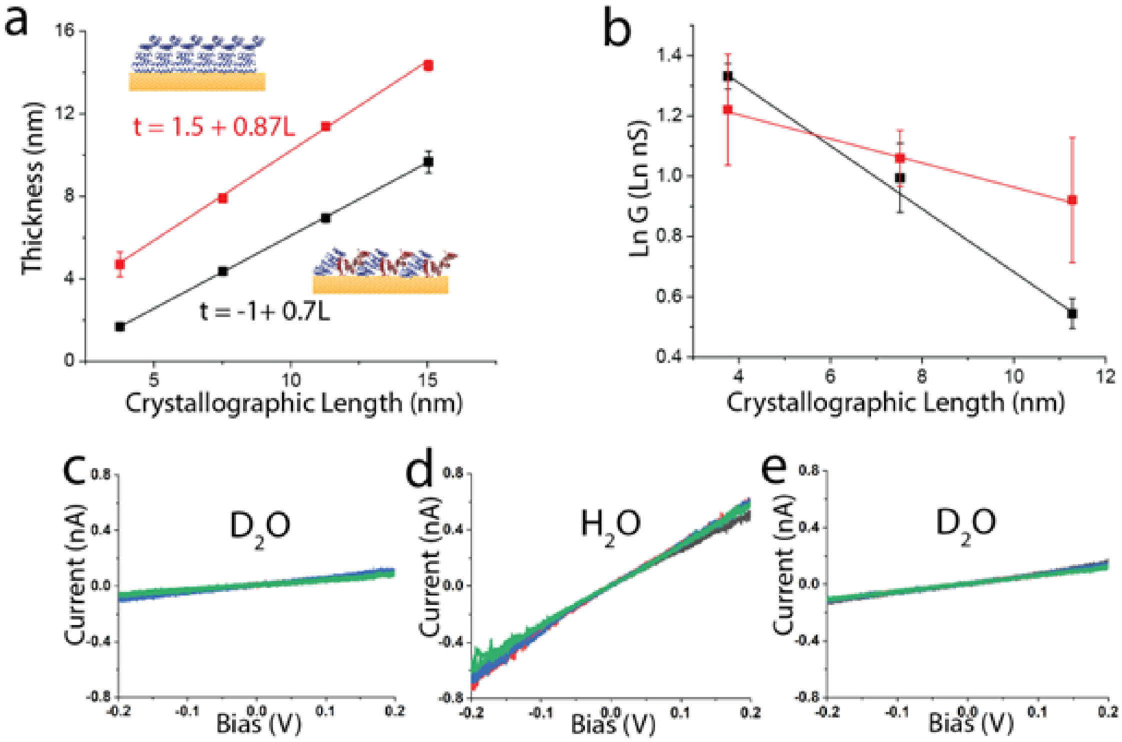


Figure 2.7: CTPR packing and $D_2O$ sensitivity. (a) Measured film thickness vs crystallographic length for CTPR4, 8, 12 and 16 deposited from 20,50,30 and 100 $\mu$M concentrations respectively (red) and 1 $\mu$M (black) with linear fits as shown. The negative intercept for the less densely packed film can be accounted for by a 2 nm shrinkage in the less dense film relative to the crystal structure. b) Natural log of conductance vs length for CTPR4, 8 and 12 in dense (red) and less dense films (black). c,d,e) Typical current vs voltage scans from the same sample of CTPR12 showing the reversible increase in conductance as the solvent is changed from $D_2O$ to $H_2O$ and back to $D_2O$.

## 2.3 Results and Discussion

The resulting unfiltered distributions of conductances (Section 2.2.3) were well-fitted by two log-normal distributions. However, as shown here, artifactual contributions from instrumental drift can dominate the lower conductance feature. The second peak, at about ten times the conductance of the first, arises from the formation of a molecular bridge between the two electrodes by two specific contacts to the protein. This peak reproducibly reflects the protein structure and the chemistry of the contacts to it.

### *2.3.1 Solvent Effects: CTPR Proteins*

In earlier work, the measured thickness of the CTPR films deposited from 1 $\mu$M solutions was only about 70% of the crystallographic length of the molecules, an effect that was attributed to the formation of a tilted layer.[19] For the present study, the effects of deposition at higher concentrations of CTPR protein was investigated. Fig. 2.7 (a) shows the monolayer thickness measured by ellipsometry for the series CTPR4, CTPR8, CTPR12, and CTPR16, plotted vs. crystallographic length as deposited overnight from 20, 50, 30 and 100 $\mu$M solutions, respectively (red points) as compared to 1 $\mu$M for all (black points) solutions. Monolayers deposited from the lower concentration (1 $\mu$M ) are about 70% of the crystallographic length. In contrast, the measured film thickness at the higher concentrations is nearly 90% of the crystallographic length, implying that the molecules are more nearly upright. The intercept for the higher concentrations (1.5 nm – fits are shown in Fig. 2.7 (a)) corresponds quite well to the length of the terminal sequence that is not part of the repeated motif. In addition to a smaller gradient, the fit to the data for films deposited at lower concentration has an unphysical negative intercept. This negative intercept may be accounted for if the molecules in the less dense film are less extended than in the crystal structure, where packing forces may serve to linearize the protein. If the exact molecular lengths are some 2 nm less, the resulting intercept would be a positive 0.5 nm.

An unexpectedly significant effect of this packing difference on the conductivity of the molecules was observed. Fig. 2.7 (b) shows the measured conductance vs. molecular length for the dense (red) and sparse films (black). The conductance is similar for the CTPR4 in both types of film but falls more rapidly with length in the less dense film. These results suggest that the packing does not influence the contact resistance because the conductance of short molecules is similar. The more rapid fall-off in conductance in the less dense film signals a smaller diffusion constant for carriers. Since the less dense film is likely more fully hydrated, this finding is consistent with the observation that increased hydration increases reorganization energy [29] which, in turn, will decrease the carrier diffusion constant.[30] In what follows, the focus is on the more densely-packed films.

A significant drop in conductance was observed on replacing the 1 mM phosphate buffer (pH 7.4) with a $D_2O$ based buffer. This effect reversed entirely on returning to the $H_2O$ buffer. Fig. 2.7 (c, d, and e) show characteristic IV curves that were taken from a sample of CTPR12 in $D_2O$ (c), then after flowing $H_2O$ into the sample (d) and again after the electrolyte was replaced with $D_2O$ (e). (About an hour was required to stabilize the microscope after these changes, limiting the time resolution of these measurements.) There are substantial molecule-to-molecule variations, and a representative sample of the data from CTPR4, 8, and 12 is given in Fig. 2.7.

Peak conductance values were obtained from repeat experiments, as listed in Table 2.2. The data have been auto-filtered as described before in Section 2.2.4 , and the data points (gray from bias sweeps up and red from bias sweeps down) are the bin cen-

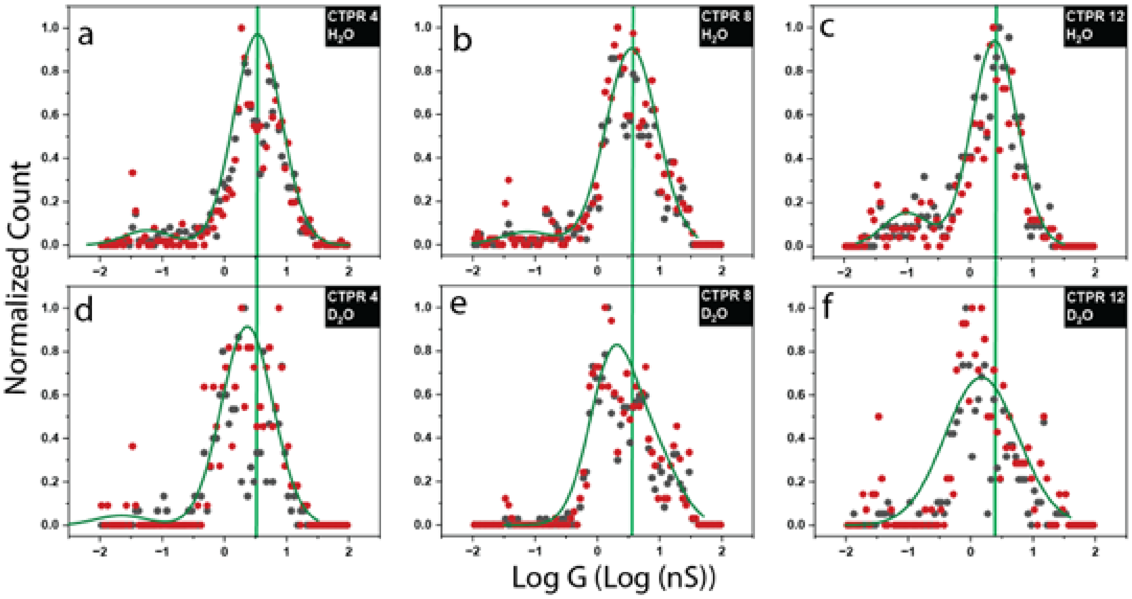


Figure 2.8: Conductance distributions in $H_2O$ and $D_2O$: Points are the bin-centers of the histogram generated from slopes of auto-filtered IV curves for sweeping up (gray) and sweeping down (red). The green curves are estimated probability density function of the unbinned auto-filtered data using a Gaussian Mixture Model (GMM) . CTPR4 in $H_2O$ (a) and $D_2O$ (d), CTPR8 in $H_2O$ (b) and $D_2O$ (e), CTPR12 in $H_2O$ (c) and $D_2O$ (f). All runs for the various samples are summarized in Table 2.2.

ters of the histograms (bin width = 0.05) formed from this filtered data. In order to locate the peaks without the artifacts caused by choice of bin size in the histograms, an unsupervised soft clustering method, namely, Gaussian Mixture Model (GMM) analysis is applied to fit the filtered (unbinned) data using a Bayesian approach.[31] These probability density function values of the GMM fits for each filtered dataset are plotted (on the same scale) as the green lines in the figures. The peaks of the distributions are shifted to smaller values of conductance (Fig. 2.8 d, e, and f) in $D_2O$ compared to $H_2O$ (Fig. 2.8 a, b, and c). The conductance scale is logarithmic, so the changes in conductance are quite significant (Table 2.1). While a single Gaussian component in the GMM is quite good, there is always some degree of residual $H_2O$ contamination for samples in $D_2O$ that were previously solvated with $H_2O$, and this can give rise to some higher conductance points (c.f. Fig. 2.8 (e)). The measurements were not extended to CTPR16 because, in $D_2O$, the higher conductance peak was not readily separated from the spurious first peak in the distribution.

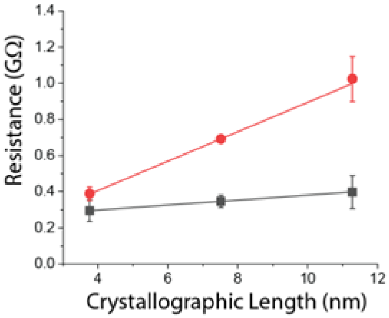


Figure 2.9: Resistance vs length for $D_2O$ (red) and $H_2O$ (black). Solid lines are fits to the linear dependence expected for hopping conductance.

The peak conductances, $\sigma_p$ for the three molecules in the two electrolytes have been converted to resistances ($\frac{1}{\sigma_p}$) and are plotted in Fig. 2.9 (black – $H_2O$, red $D_2O$) and fitted to

$$R(L) = R_0 + \rho \frac{L}{A}, \tag{2.1}$$

where $R_0$ is the contact resistance and $\frac{\rho}{A}$ is the ratio of resistivity to the effective cross-sectional area of the molecule, modeled as a conducting rod. It is seen that the contact resistance is not changed above experimental uncertainty on deuteration, with $R_0(H_2O) = 0.24 \pm 0.1$ GΩ and $R_0(D_2O) = 0.10 \pm 0.07$ GΩ. The slopes differ substantially with $\frac{\rho}{A}(H_2O) = 0.013 \pm 0.001$ and $\frac{\rho}{A}(D_2O) = 0.074 \pm 0.005$.

Substitution of $D_2O$ has, at most, a negligible effect on the contact resistance (which reflects the barrier to charge injection [32]). However, the slope of the resistance vs. molecular length plot (Fig. 2.9) changes significantly, indicating a 5.8-fold increase in resistivity on deuteration. The carrier diffusion constants are inversely

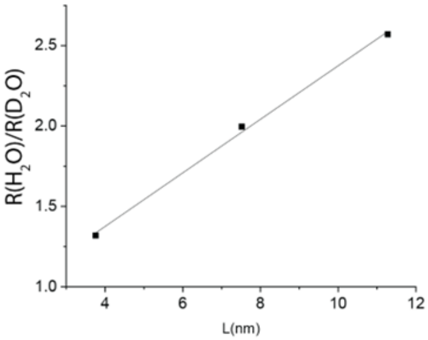


Figure 2.10: Ratio of resistances in $H_2O$ to those in $D_2O$ as a function of length for the CTPR proteins. The solid line is a fit with intercept 0.68 ± 0.73 and slope 0.174 ± 0.1/nm.

related to resistivity, so diffusion in the $H_2O$ solvated protein is  6 times as rapid as in the $D_2O$ solvated protein. The ratio of the resistances in $D_2O$ and $H_2O$ increases by 0.174 ± 0.1 /nm (Fig. 2.10), so, for the 460 nm OmcS wires measured by Dahl et al. [21], this dependence would predict  80-fold increase in resistance, which is about 2.5x less than the 200 fold change observed. Nonetheless, this extrapolation is within an order of magnitude of the observed result, suggesting that the mechanism is similar in both cases. The more significant effect in OmcS may reflect the greater number of exchangeable protons on the hemes [33] in the OmcS proteins.

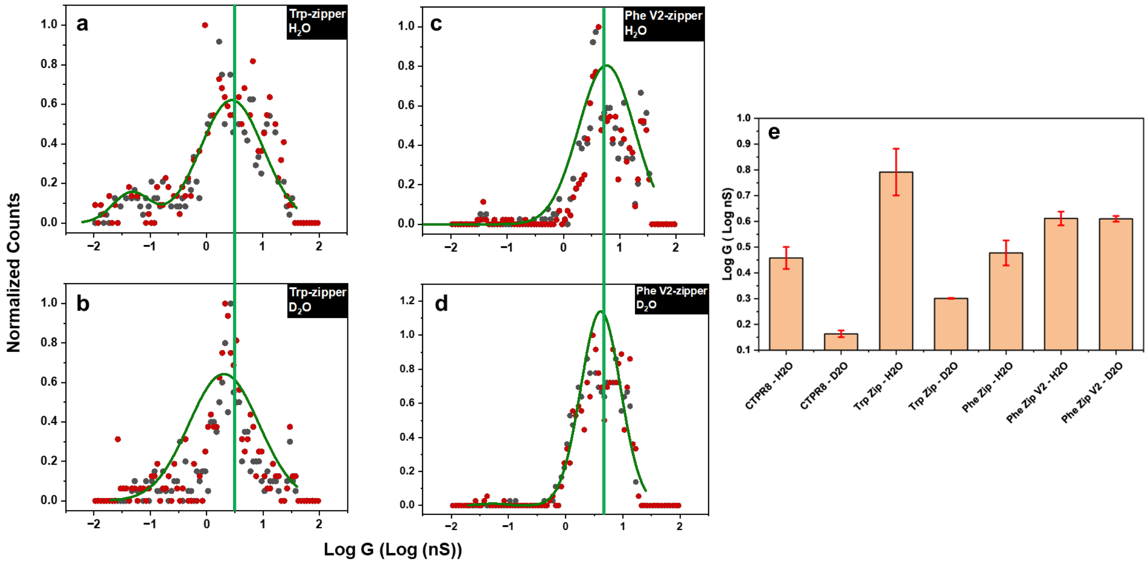


Figure 2.11: Tryptophan Zipper Proteins Are Sensitive to Deuteration, Phenylalanine Zippers Are Not. Distribution of conductances for the Trp-zipper in a) $H_2O$ and b) $D_2O$ and the Phe V2-zipper in c) $H_2O$ and d) $D_2O$. e) Comparison of peak values for log conductance for CTPR 8, the Trp-zipper and the Phe-zipper in $D_2O$ and $H_2O$. Enhanced conductance in the zipper proteins is accounted for by the closer packing of aromatic residues. The Phe-zipper is insensitive to deuteration.

| **Molecule** | **G ($H_2O$) nS** | **G ($D_2O$) nS** |
|---|---|---|
| CTPR4 | 3.39±0.69 | 2.57±0.25 |
| CTPR8 | 2.88±0.28 | 1.45±0.03 |
| CTPR12 | 2.51±0.58 | 0.98±0.12 |
| Phe Zipper V2 | 4.08±0.26 | 3.97±0.03 |
| Trp Zipper | 5.03±0.26 | 1.99±0.01 |

Table 2.1: Molecular Conductances in $H_2O$ and $D_2O$. Averages and Error are from Three Repeats.

### *2.3.2 Solvent Effects: Zipper Proteins*

More insight is gained from a study of the two types of zipper proteins. The phenylalanine zipper structure (2GUV) is slightly longer than the tryptophan zipper, so, a mutant with a 3 aa deletion, labeled Phe V2, in the Fig. 2.3 (b) (sequences are shown in the Fig. 2.3) was also expressed. Panels (a) and (b) in Fig. 2.11 show the histogrammed data for the Trp-Zipper ((a) and (b) for $H_2O$ and $D_2O$, respectively) together with the green curves that are the GMM probability estimation modeled on the raw (unbinned) data. Data and GMM fits for the Phe V2 zipper are shown in panels (c) and (d) in Fig. 2.11(data for the original Phe zipper in $H_2O$ are listed in Table 2.3). As summarized in Fig. 2.11 (e) and Table 2.1, the Trp-Zipper shows a similar change to that of the CTPR8 wires (also about 7 nm in length) on deuteration. However, the Phe V2-Zippers do not show a significant change. The change shown by the Trp-Zipper suggests that the indole proton in Trp is the site that is specifically modified in $D_2O$.

The lack of any significant effect of deuteration on the Phe-Zipper shows that the effect is specific to the (exchangeable) proton on the indole nitrogen in the case of tyrosine. Three changes are possible: Firstly, the proton is exchanged for a deuteron.

Secondly, D...ND bonding occurs, and thirdly, the proton is not exchanged rapidly, and D...NH bonding occurs. The researchers note that Cioni and Strambini [9] have found that the effects of deuteration on protein flexibility are entirely in place before H-D exchange occurs, so D...NH bonding may cause rapid changes, although, rescanning samples was possible only after an hour of equilibration. Changes in protein flexibility on deuteration are presumably similar for the three proteins studied here, so the negative result in the case of the Phe-Zipper suggests that flexibility changes are not the major contributor to $D_2O$ sensitivity. The data collection that is shown in Fig. 2.11 (e) and Table 2.1 has other implications. Since the phenylalanine residues cannot form stable radical cations, it might seem that hopping conductivity could not occur in this molecule. However, Shapiro et al. [34] have shown that inserting phenylalanine residues into bacterial filaments enhances their conductivity. The presented measurements show that the Phe-zipper is nearly as conductive as the Trp-Zipper, for which the aromatic residues are readily oxidized. A molecular simulation of the CTPR proteins [32] shows that the charged states of the aromatic residues are far from fully relaxed, with charge transfer occurring on 100 ps timescales compared to $\mu$s time for full relaxation of the environment. Thus, the carriers may correspond to a transiently charged state, in which the reorganization energy is a fraction of that possible for a fully relaxed radical cation (were such a state possible for the Phe residues). The original zipper (Table 2.3) was only 75% as conductive as the second version with a 3aa deletion. The additional length alone cannot account for this change because of the slow decay of current with distance. The difference presumably reflects differences in the barrier to charge injection, where tunneling transport dominates, leading to a strong distance dependence in the contact region. [32] Another interesting feature is that the conductance of the zippers is only about 1.5x higher than that of the CTPR8, despite the closer packing of aromatic residues in the zipper proteins. However, a simple calculation shows that this is not unexpected. The average edge-to-edge distances of aromatic nearest-neighbor residue pairs in CTPR8 is 0.72 nm, and for the Trp-Zipper, the average is 0.37 nm. Taking the electronic decay constant to be 27 0.77 nm-1, it is estimated that the Trp-Zipper should be 1.6x as conductive as the CTPR8, similar to the factor 1.42 observed (Table 2.1). In the case of multiheme proteins, a model has been proposed in which a valence band, made from many residue contributions, sustains direct tunneling across the protein.[35] The dramatic change in solvent response of the zipper proteins on changing the aromatic residue supports the proposal that these aromatic residues dominate transport in the non-redox active proteins studied in this chapter. In that case, transport is sustained by a long-range hopping mechanism with a small reorganization energy.[32]

| **Molecule** | **Solvent** | **Run** | **Gap** | **Log G Manual** | **Log G Auto** |
|---|---|---|---|---|---|
| | | | **nm** | **Log (nS)** | **Log (nS)** |
| **CTPR4** | $H_2O$ | 1 | Z0 + 1 nm | 0.76 | 0.68 |
| | $H_2O$ | 2 | Z0 + 1 nm | 0.52 | 0.49 |
| | $H_2O$ | 3 | Z0 + 1 nm | 0.50 | 0.42 |
| **Avg (Log G)** | | | | **0.59 ± 0.14** | **0.53 ± 0.13** |
| **CTPR4** | $D_2O$ | 1 | Z0 + 1 nm | 0.25 | 0.44 |
| | $D_2O$ | 2 | Z0 + 1 nm | 0.31 | 0.33 |
| | $D_2O$ | 3 | Z0 + 1 nm | 0.21 | 0.45 |
| **Avg (Log G)** | | | | **0.26 ± 0.05*** | **0.41 ± 0.07** |
| **CTPR8** | $H_2O$ | 1 | Z0 + 3 nm | 0.52 | 0.51 |
| | $H_2O$ | 2 | Z0 + 3 nm | 0.51 | 0.48 |
| | $H_2O$ | 3 | Z0 + 3 nm | 0.38 | 0.38 |
| **Avg (Log G)** | | | | **0.47 ± 0.08** | **0.46 ± 0.07** |
| **CTPR8** | $D_2O$ | 1 | Z0 + 3 nm | -0.04 | 0.19 |
| | $D_2O$ | 2 | Z0 + 3 nm | -0.02 | 0.16 |
| | $D_2O$ | 3 | Z0 + 3 nm | -0.20 | 0.14 |
| **Avg (Log G)** | | | | **0.09 ± 0.1** | **0.16 ± 0.03** |
| **CTPR 12** | $H_2O$ | 1 | Z0 + 4 nm | 0.48 | 0.47 |
| | $H_2O$ | 2 | Z0 + 4 nm | 0.49 | 0.51 |
| | $H_2O$ | 3 | Z0 + 4 nm | 0.41 | 0.23 |
| **Avg (Log G)** | | | | **0.46 ± 0.04** | **0.4 ± 0.15** |
| **CTPR 12** | $D_2O$ | 1 | Z0 + 4 nm | -0.24 | 0.09 |
| | $D_2O$ | 2 | Z0 + 4 nm | -0.19 | -0.09 |
| | $D_2O$ | 3 | Z0 + 4 nm | -0.21 | -0.03 |
| **Avg (Log G)** | | | | **-0.21 ± 0.03** | **-0.01 ± 0.09** |
| **CTPR16** | $H_2O$ | 1 | Z0 + 5 nm | 0.37 | 0.26 |

Table 2.2: CTPR Proteins: Summary of conditions and results for each run. The "Manual" columns list the peak conductances obtained from manual filtering and Gaussian fits to histograms. The "Auto" column lists values obtained from the Python script-based automated analysis described in section 2.2.4. The two methods are in agreement, with the exception of the data marked with an asterisk where the difference is >1SD.

| **Molecule** | **Solvent** | **Run** | **Gap** | **Log G Manual** | **Log G Auto** |
|---|---|---|---|---|---|
| | | | **nm** | **Log (nS)** | **Log (nS)** |
| **Phe Zipper** | $H_2O$ | 1 | Z0 + 2 nm | 0.7 | 0.526 |
| | $H_2O$ | 2 | Z0 + 2 nm | 0.85 | 0.429 |
| **Avg (Log G)** | | | | **0.78 ± 0.11*** | **0.48 ± 0.07** |
| **Phe Zipper V2** | $H_2O$ | 1 | Z0 + 0 nm | 0.56 | 0.658 |
| | $H_2O$ | 2 | Z0 + 0 nm | 0.53 | 0.565 |
| | $H_2O$ | 3 | Z0 + 2 nm | 0.75 | 0.611 |
| **Avg (Log G)** | | | | **0.61 ± 0.07** | **0.611 ± 0.05** |
| **Phe Zipper V2** | $D_2O$ | 1 | Z0 + 2 nm | 0.74 | 0.60 |
| | $D_2O$ | 2 | Z0 + 2 nm | 0.62 | 0.60 |
| | $D_2O$ | 3 | Z0 + 2 nm | 0.55 | 0.63 |
| **Avg (Log G)** | | | | **0.64 ± 0.05** | **0.61 ± 0.02** |
| **Trp Zipper** | $H_2O$ | 1 | Z0 + 1 nm | 0.83 | 0.680 |
| | $H_2O$ | 2 | Z0 + 2 nm | 0.92 | 0.971 |
| | $H_2O$ | 3 | Z0 + 1 nm | 0.79 | 0.724 |
| **Avg (Log G)** | | | | **0.84 ± 0.04** | **0.79 ± 0.16** |
| **Trp Zipper** | $D_2O$ | 1 | Z0 + 1 nm | 0.35 | 0.30 |
| | $D_2O$ | 2 | Z0 + 2 nm | 0.3 | 0.30 |
| | $D_2O$ | 3 | Z0 + 1 nm | 0.22 | 0.30 |
| **Avg (Log G)** | | | | **0.29 ± 0.07** | **0.30** |

Table 2.3: Zipper Proteins: Summary of conditions and results for each run. The “Manual” columns list the peak conductances obtained from manual filtering and Gaussian fits to histograms. The “Auto” column lists values obtained from the Python script-based automated analysis described in section 2.2.4. The two methods are in agreement, with the exception of the data marked with an asterisk where the difference is >1SD.

## 2.4 Conclusion

In summary, this chapter has shown that the reduction of molecular conductance on deuteration reported earlier for OmcS bacterial filaments is also observed at a similar magnitude (on a unit length basis) in proteins that contain no redox co-factors. The lack of any effect of deuteration in the case of a protein in which the aromatic residues contain no exchangeable protons (Phe-Zipper) demonstrates that the dominant effect is the exchange of the binding of a deuteron at exchangeable sites on the aromatic residues of the protein. We have also shown how an automated analysis can remove artifactual features from the conductance distributions. Furthermore, we have highlighted how Gaussian Mixture Modelling (a Bayesian approach), unlike histograms, can provide a more accurate estimate of the underlying distribution parameters and is not constrained by the bin size.

# REFERENCES

[1] Sibel Ebru Yalcin and Nikhil S. Malvankar. The blind men and the filament: Understanding structures and functions of microbial nanowires. *Current Opinion in Chemical Biology*, 59:193–201, 2020.

[2] Christopher D Bostick, Sabyasachi Mukhopadhyay, Israel Pecht, Mordechai Sheves, David Cahen, and David Lederman. Protein bioelectronics: a review of what we do and do not know. *Reports on Progress in Physics*, 81(2):026601, 2018.

[3] Nicole L. Ing, Mohamed Y. El-Naggar, and Allon I. Hochbaum. Going the Distance: Long-Range Conductivity in Protein and Peptide Bioelectronic Materials. *The Journal of Physical Chemistry B*, 122(46):10403–10423, 2018.

[4] Tracy Q. Ha, Inco J. Planje, Jhanelle R.G. White, Albert C. Aragonès, and Ismael Díez-Pérez. Charge transport at the protein–electrode interface in the emerging field of BioMolecular Electronics. *Current Opinion in Electrochemistry*, 28:100734, 2021.

[5] Bintian Zhang, Weisi Song, Pei Pang, Huafang Lai, Qiang Chen, Peiming Zhang, and Stuart Lindsay. Role of contacts in long-range protein conductance. *Proceedings of the National Academy of Sciences*, 116(13):5886–5891, 2019.

[6] Stuart Lindsay. Ubiquitous Electron Transport in Non-Electron Transfer Proteins. *Life*, 10(5), 2020.

[7] Dmitry V. Matyushov. Solvent reorganization energy of electron-transfer reactions in polar solvents. *The Journal of Chemical Physics*, 120(16):7532–7556, 2004.

[8] D J Kushner, Alison Baker, and T G Dunstall. Pharmacological uses and perspectives of heavy water and deuterated compounds. *Canadian Journal of Physiology and Pharmacology*, 77(2):79–88, 1999.

[9] Patrizia Cioni and Giovanni B. Strambini. Effect of Heavy Water on Protein Flexibility. *Biophysical Journal*, 82(6):3246–3253, 2002.

[10] Judith P. Klinman and Amnon Kohen. Hydrogen Tunneling Links Protein Dynamics to Enzyme Catalysis. *Annual Review of Biochemistry*, 82(1):471–496, 2013.

[11] Michael J Weaver and Scott M Nettles. Solvent isotope effects upon the thermodynamics of some transition-metal redox couples in aqueous media. *Inorganic Chemistry*, 19(6):1641–1646, 1980.

[12] Ole Farver, Jingdong Zhang, Qijin Chi, Israel Pecht, and Jens Ulstrup. Deuterium isotope effect on the intramolecular electron transfer in *Pseudomonas aeruginosa* azurin. *Proceedings of the National Academy of Sciences*, 98(8):4426–4430, 2001.

[13] Daniel H. Murgida and Peter Hildebrandt. Proton-Coupled Electron Transfer of Cytochrome c. *Journal of the American Chemical Society*, 123(17):4062–4068, 2001.

[14] Martin Byrdin, Valérie Sartor, André P.M. Eker, Marten H. Vos, Corinne Aubert, Klaus Brettel, and Paul Mathis. Intraprotein electron transfer and proton dynamics during photoactivation of DNA photolyase from E. coli: review and new insights from an “inverse” deuterium isotope effect. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1655:64–70, 2004.

[15] Setare Mostajabi Sarhangi and Dmitry V. Matyushov. Effect of Water Deuteration on Protein Electron Transfer. *The Journal of Physical Chemistry Letters*, 14(3):723–729, 2023.

[16] Juan M. Artés, Ismael Díez-Pérez, Fausto Sanz, and Pau Gorostiza. Direct Measurement of Electron Transfer Distance Decay Constants of Single Redox Proteins by Electrochemical Tunneling Spectroscopy. *ACS Nano*, 5(3):2060–2066, 2011.

[17] Marta P. Ruiz, Albert C. Aragonès, Nuria Camarero, J. G. Vilhena, Maria Ortega, Linda A. Zotti, Rubén Pérez, Juan Carlos Cuevas, Pau Gorostiza, and Ismael Díez-Pérez. Bioengineering a Single-Protein Junction. *Journal of the American Chemical Society*, 139(43):15337–15346, 2017.

[18] Anna Lagunas, Alejandra Guerra-Castellano, Alba Nin-Hill, Irene Díaz-Moreno, Miguel A De la Rosa, Josep Samitier, Carme Rovira, and Pau Gorostiza. Long distance electron transfer through the aqueous solution between redox partner proteins. *Nature Communications*, 9(1):5157, 2018.

[19] Bintian Zhang, Eathen Ryan, Xu Wang, Weisi Song, and Stuart Lindsay. Electronic Transport in Molecular Wires of Precisely Controlled Length Built from Modular Proteins. *ACS Nano*, 16(1):1671–1680, 2022.

[20] Ramesh Y Adhikari, Nikhil S Malvankar, Mark T Tuominen, and Derek R Lovley. Conductivity of individual Geobacter pili. *RSC Advances*, 6(10):8354–8357, 2016.

[21] Peter J. Dahl, Sophia M. Yi, Yangqi Gu, Atanu Acharya, Catharine Shipps, Jens Neu, J. Patrick O'Brien, Uriel N. Morzan, Subhajyoti Chaudhuri, Matthew J. Guberman-Pfeffer, Dennis Vu, Sibel Ebru Yalcin, Victor S. Batista, and Nikhil S. Malvankar. A 300-fold conductivity increase in microbial cytochrome nanowires

due to temperature-induced restructuring of hydrogen bonding networks. *Science Advances*, 8(19):eabm7193, 2022.

[22] Anthony Harriman. Further comments on the redox potentials of tryptophan and tyrosine. *Journal of Physical Chemistry*, 91(24):6102–6104, 1987.

[23] Michael Tuchband, Jin He, Shuo Huang, and Stuart Lindsay. Insulated gold scanning tunneling microscopy probes for recognition tunneling in an aqueous environment. *Review of Scientific Instruments*, 83(1):015102, 2012.

[24] Tommi Kajander, Aitziber L Cortajarena, Simon Mochrie, and Lynne Regan. Structure and stability of designed TPR protein superhelices: unusual crystal packing and implications for natural TPR proteins. *Acta Crystallographica Section D: Biological Crystallography*, 63(7):800–811, 2007.

[25] Sara H Mejías, Begoña Sot, Raul Guantes, and Aitziber L Cortajarena. Controlled nanometric fibers of self-assembled designed protein scaffolds. *Nanoscale*, 6(19):10982–10988, 2014.

[26] Sara H Mejias, Pierre Couleaud, Santiago Casado, Daniel Granados, Miguel Angel Garcia, Jose M Abad, and Aitziber L Cortajarena. Assembly of designed protein scaffolds into monolayers for nanoparticle patterning. *Colloids and Surfaces B: Biointerfaces*, 141:93–101, 2016.

[27] Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7):1731–1737, 2003.

[28] Shuai Chang, Jin He, Peiming Zhang, Brett Gyarfas, and Stuart Lindsay. Gap

Distance and Interactions in a Molecular Tunnel Junction. *Journal of the American Chemical Society*, 133(36):14267–14269, 2011.

[29] Edward L. Mertz and Lev I. Krishtalik. Low dielectric response in enzyme active site. *Proceedings of the National Academy of Sciences*, 97(5):2081–2086, 2000.

[30] Yoni Eshel, Uri Peskin, and Nadav Amdursky. Coherence-assisted electron diffusion across the multi-heme protein-based bacterial nanowire. *Nanotechnology*, 31(31):314002, 2020.

[31] Jake VanderPlas. *Python data science handbook: Essential tools for working with data*. O'Reilly Media, Inc., 2016.

[32] Siddharth Krishnan, Aleksei Aksimentiev, Stuart Lindsay, and Dmitry Matyushov. Long-range Hopping Conductivity in Proteins. *bioRxiv*, pages 2022–10, 2022.

[33] Juliette TJ Lecomte and Gerd N La Mar. Proton NMR study of labile proton exchange in the heme cavity as a probe for the potential ligand entry channel in myoglobin. *Biochemistry*, 24(25):7388–7395, 1985.

[34] Daniel Mark Shapiro, Gunasheil Mandava, Sibel Ebru Yalcin, Pol Arranz-Gibert, Peter J Dahl, Catharine Shipps, Yangqi Gu, Vishok Srikanth, Aldo I Salazar-Morales, J Patrick O'Brien, et al. Protein nanowires with tunable functionality and programmable self-assembly using sequence-controlled synthesis. *Nature Communications*, 13(1):829, 2022.

[35] Zdenek Futera, Ichiro Ide, Ben Kayser, Kavita Garg, Xiuyun Jiang, Jessica H. van Wonderen, Julea N. Butt, Hisao Ishii, Israel Pecht, Mordechai Sheves, David Cahen, and Jochen Blumberger. Coherent Electron Transport across a 3 nm

Bioelectronic Junction Made of Multi-Heme Proteins. *The Journal of Physical Chemistry Letters*, 11(22):9766–9774, 2020.

# Chapter 3

# DECODING PROTEIN IMPACT ON CONDUCTANCE MEASUREMENTS USING MACHINE LEARNING

*In this chapter, the solid-state chip design and fabrication were done by Dr. Joshua Sadar and Dr. Weisi Song. The current vs. time measurements were carried out by Dr. Weisi Song. My contribution included the machine learning algorithm design and the statistical analysis of the experimental data.*

## 3.1 Introduction

The ability to monitor single-molecule conductance is critical for understanding electron transport at the molecular level and developing advanced bio-sensing technologies. The aim of this work is to enhance the scalability of the measurements by transitioning from STM to solid-state chips. However, it has been observed that the chip signals are more complex compared to those from the STM at this point. As a result, a machine learning approach is imperative to initiate their interpretation. By integrating single-molecule electrical measurement with machine learning, researchers achieved high-precision identification of biomolecules [1, 2] , previously difficult using conventional methods. This chapter builds upon the findings of the previous chapter by creating a machine-learning model to detect the presence of a specific protein and to identify the crucial features necessary for the detection process. In order to achieve this, a sophisticated solid-state device was designed, which is composed of two metallic layers separated by an insulating layer approximately 6 nm thick. This insulating layer was deposited using an advanced technique called plasma-enhanced atomic layer deposition (PEALD), which offers precise control over the layer's thickness and composition. By employing this device, the limitations associated with characterizing the scanning tunneling microscopy (STM) gap can be addressed, resulting in the acquisition of more stable current versus time data and a better signal-to-noise ratio.

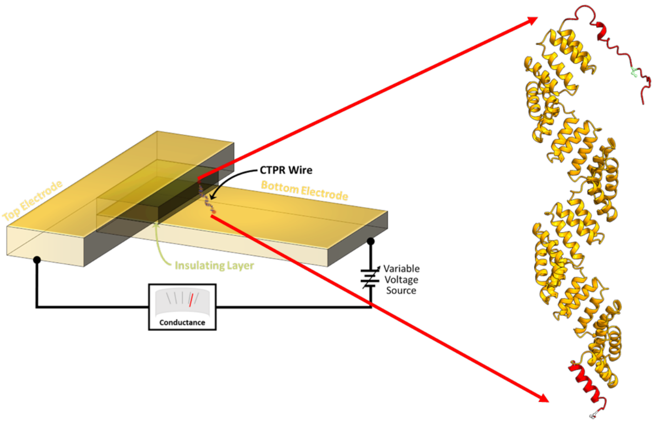


Figure 3.1: Schematic of Basic Solid-State Device-aided Junction to Contact CTPR Protein Wires. The junction gap is set by the thickness of the PEALD insulating layer, which permits a tunable and controllable junction gap to be targeted. The CTPR16 protein wire is zoomed-in showing cysteine residues used for modification shown in white in the N- and C- termini. (*Schematic of the chip created by Dr. Joshua Sadar*)

Notably, once this technology is operational, it will be scalable, allowing for the formation of numerous protein junctions on a single chip. Through the analysis of the electronic current signatures, this study aims to gain a preliminary understanding of the changes that occur when a protein bridges the gap between the two metallic layers in the solid-state device. By utilizing this information, a machine-learning model can be designed to identify the specific patterns and features that are associated with the presence of CTPR16 protein wires in the junction with improved accuracy and performance. Hence, this work has the potential to contribute significantly to the field of protein detection and characterization in solid-state chips, ultimately advancing our understanding of protein functions and interactions at the molecular level.

## 3.2 Experimental Setup

A solid-state wafer with 32 channel chip was used to monitor the long-range conductance of proteins due to electron transport. Out of the 32 devices, we received signals from seven devices. The schematic in Fig. 3.1 shows the basic architecture of the wafer. The details of the fabrication process are beyond the scope of this thesis. In short, the electrodes were covered with a passivation layer that was etched open to expose the junction area. Junctions were protected from electrostatic damage with aluminum shorts that were etched away immediately prior to use. The next step involved timetch etching of the insulating layer to remove the insulating layer from over the bottom electrode (retaining it inside the gap to leave a structure like that shown in Figure 3.1) and wire bonding the chip to a chip carrier. Next, a 1.2 cm x 1.2 cm PDMS cube was prepared, and a 7 mm hole was drilled through the cube in the center. A sonicated (1:1 ethanol to Nanopure water), cleaned and dried Polydimethylsiloxane (PDMS) sample cell was mounted on top of the cleaned chip using Kwik-Sil adhesive.Filtered 1 mM Phosphate Buffer (PB buffer) was added to the sample cell, and current vs time data was recorded with an 8-channel Data Acquisition (DAQ) system at 50 mV bias. This data will be referred to as 'Control' data. The device was then functionalized with 1 $\mu$M cysteine-terminated CTPR16 (Fig. 3.1) solution in freshly made PB buffer for two hours. The data collected at this experimental condition will be referred to as 'CTPR16' in the next sections.

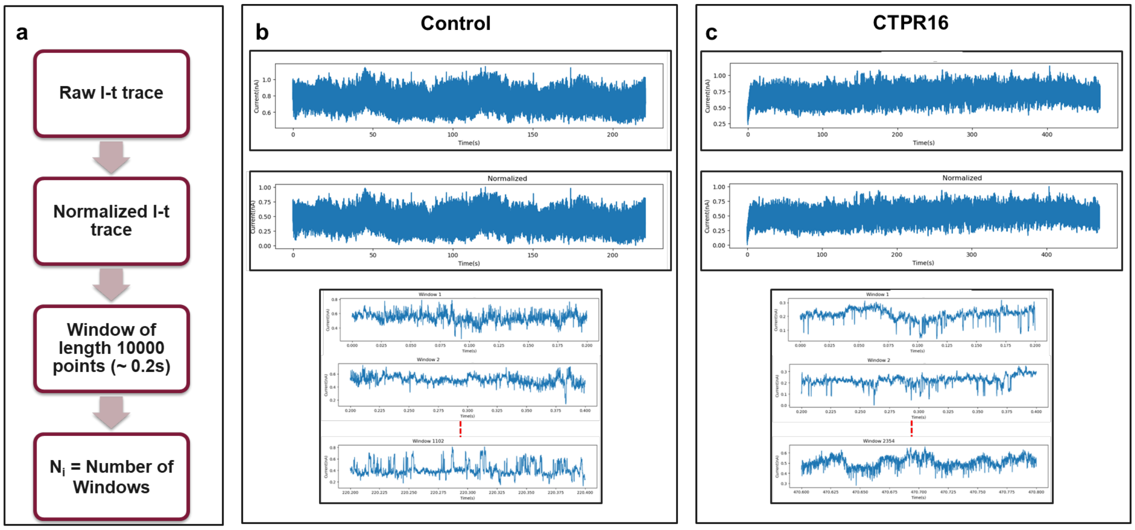
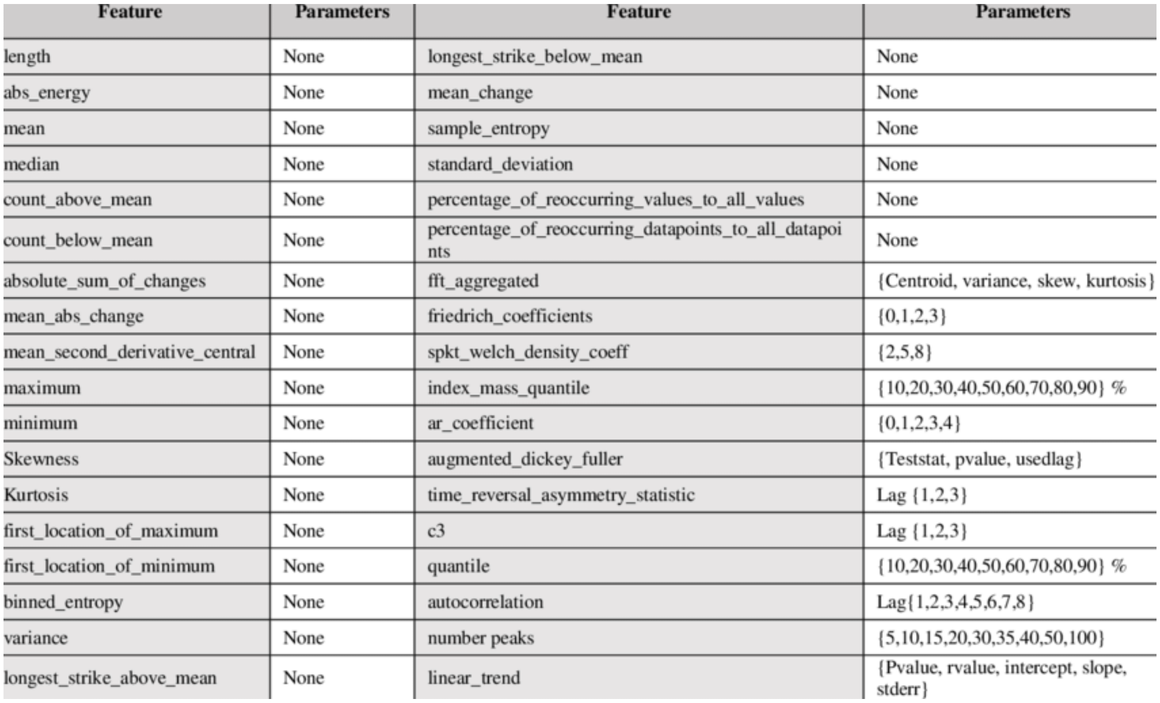


Figure 3.2: Pipeline for Pre-processing the Current vs. Time Data. (a) The steps include normalization and windowing. (b) Pre-processing steps for CONTROL data from a sample device. The top-most graph in this panel is the entire raw current vs. time trace for a given device, followed by the normalized current vs. time trace for the same device and the bottom series of graphs are the result of breaking the etire normalized current vs time trace into windows of 10000 points each. (c) Pre-processing steps for CTPR16 data from the same sample device. The top-most graph in this panel is the entire raw current vs. time trace for a given device, followed by the normalized current vs. time trace for the same device and the bottom series of graphs are the result of breaking the etire normalized current vs time trace into windows of 10000 points each.

## 3.3 Feature Engineering

In this section, the proposed feature engineering process is investigated. Feature engineering pertains to extracting and selecting the most significant features from a given dataset, typically executed to enhance the learning capabilities of a machine learning (ML) model. The pre-requisite for feature extraction and selection is data pre-processing. It entails cleaning, transforming, and organizing raw data into a structured format amenable to further analysis. In this study, the data pre-processing techniques of normalization and windowing were applied to the entire current versus time trace, preparing the dataset for feature extraction and machine learning. Normalization is a critical data pre-processing approach that scales data values within a specific range, typically [0, 1], to eliminate biases stemming from differences in data value scales across all the devices. In this study, the current vs. time trace from every device was normalized to ensure comparability among all data points and to prevent variations in the magnitudes of the data points from influencing the subsequent analysis. Normalization contributes to the reduction of outlier impacts ensuring that the noise in the data gets considered as outliers and does not affect the future analysis. Normalization also enhances the convergence and performance of machine learning algorithms. Following the normalization of the current vs. time trace, it was segmented into windows of 10,000 data points, with each window corresponding to a duration of 0.2 seconds, given the 50 kHz sampling frequency. Windowing is a crucial step in time-series data analysis, as it enables the exploration of localized features and patterns within the dataset (see Fig. 3.2). By partitioning the normalized current vs. time trace into smaller segments, the study concentrates on extracting valuable information from each window to be employed in subsequent feature extraction and machine learning tasks. The windowing ensures two things: generation of a large number of sample traces from each device rather than just one long trace and reduction in the computational cost. A large number of samples are needed for binary classification using deep learning to ensure accurate parameter estimation, robust feature representation, adequate representation of both classes, better handling of noise and variability, and improved optimization of the loss function.

The 'FRESH' algorithm achieves the initial task in this study by extracting features from the pre-processed dataset. For computational convenience, the authors of the FRESH algorithm have developed a standardized Python-based package called "ts-fresh", incorporating the FRESH algorithm within its framework. The source code and GitHub page of the ts-fresh package can be found in the link provided in



| Feature | Parameters | Feature | Parameters |
|---|---|---|---|
| length | None | longest_strike_below_mean | None |
| abs_energy | None | mean_change | None |
| mean | None | sample_entropy | None |
| median | None | standard_deviation | None |
| count_above_mean | None | percentage_of_reoccurring_values_to_all_values | None |
| count_below_mean | None | percentage_of_reoccurring_datapoints_to_all_datapoints | None |
| absolute_sum_of_changes | None | fft_aggregated | {Centroid, variance, skew, kurtosis} |
| mean_abs_change | None | friedrich_coefficients | {0,1,2,3} |
| mean_second_derivative_central | None | spkt_welch_density_coeff | {2,5,8} |
| maximum | None | index_mass_quantile | {10,20,30,40,50,60,70,80,90} % |
| minimum | None | ar_coefficient | {0,1,2,3,4} |
| Skewness | None | augmented_dickey_fuller | {Teststat, pvalue, usedlag} |
| Kurtosis | None | time_reversal_asymmetry_statistic | Lag {1,2,3} |
| first_location_of_maximum | None | c3 | Lag {1,2,3} |
| first_location_of_minimum | None | quantile | {10,20,30,40,50,60,70,80,90} % |
| binned_entropy | None | autocorrelation | Lag{1,2,3,4,5,6,7,8} |
| variance | None | number peaks | {5,10,15,20,30,35,40,50,100} |
| longest_strike_above_mean | None | linear_trend | {Pvalue, rvalue, intercept, slope, stderr} |

Figure 3.3: Extracted Features Using the FRESH Algorithm [3]

[4]. Nearly 800 features are automatically extracted from each current vs. time data window. Some of the extracted features are mentioned in Fig. 3.3.

An important step in reducing the computational cost is the feature selection process. During feature selection, the features in the data that contribute most to the target class (i.e., Control and CTPR16) are selected. 'SelectKBest' from the scikit-learn Python library assists in choosing the best predictors for the target class. This algorithm computes the Chi-square between each feature and the target, selecting the desired number of features with the best Chi-square scores or the lowest p-values. The Chi-squared ($\chi^2$) test is utilized in statistics to test the independence of two events. More specifically, during feature selection, it is employed to test whether the occurrence of a specific feature and the target class are independent or not. For each feature and target combination, a corresponding high $\chi^2$ score or a low p-value indicates that the target column depends on the feature column. The number of features to be selected can be user-defined. Here, 700 most important features were selected out of 800 extracted features. It should be noted that a simple Chi-squared ($\chi^2$) test assists in prioritizing features and cannot be used for future prediction because it learns linear dependencies between the feature and the target class. Hence, a more complex dependency needs to be studied by employing an optimized neural network known for efficient binary classification. This is explained in detail in Section 3.4.3.

Since the feature set extracted by the FRESH algorithm and selected by SelectKBest algorithm contains diverse data points scattered over an extensive range, features with higher magnitudes may introduce bias during model training. Thus, it is crucial to standardize the accumulated features to a standard scale. The current study employs the feature-wise Min–Max data standardization method to address this issue. Min–Max transforms features by scaling each feature to a given range of 0 to 1 using the following mathematical expression:

$$[f(x_i) = \frac{x_i - \min(X)}{\max(X) - \min(X)}] \tag{3.1}$$

where $X$ is a vector composed of $x_i$ (each feature column generated by tsfresh), and $\min(X)$ and $\max(X)$ are the minimum and maximum values of $X$, respectively. These steps result in an $N \times 700$ matrix, where $N$ is the number of 10,000-length time series windows for each device. The features for 'Control' and 'CTPR16 Protein' are vertically stacked and shuffled before feature selection and scaling. This strategically crafted dataset can now be used for training the model (Fig. 3.5).

## 3.4 Model Training

### *3.4.1 Evaluation Metrics*

In supervised machine learning techniques, labeled data is provided to the classifier for training purposes. The trained model is then evaluated for its ability to efficiently predict and generalize on unlabeled data. Model generalization refers to the ability of a trained machine learning model to perform well on new, unseen data that is not part of the training set. Suppose a model only performs well on the training data and does not generalize well. In that case, it may be overfitting to the training data, meaning it has learned to model the noise and idiosyncrasies of the training data rather than the underlying patterns. The performance of such a model (classifier) is assessed based on several performance evaluation metrics. The mathematical expressions for calculating the evaluation metrics are depicted in the following equations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3.2}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{3.3}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3.4}$$

where,

- True Positive (TP): The number of cases where the model predicted a positive result and the actual result was positive.
- True Negative (TN): The number of cases where the model predicted a negative result and the actual result was negative.
- False Positive (FP): The number of cases where the model predicted a positive result, but the actual result was negative. False positives are also known as

Type I errors.

- False Negative (FN): The number of cases where the model predicted a negative result, but the actual result was positive. False negatives are also known as Type II errors.

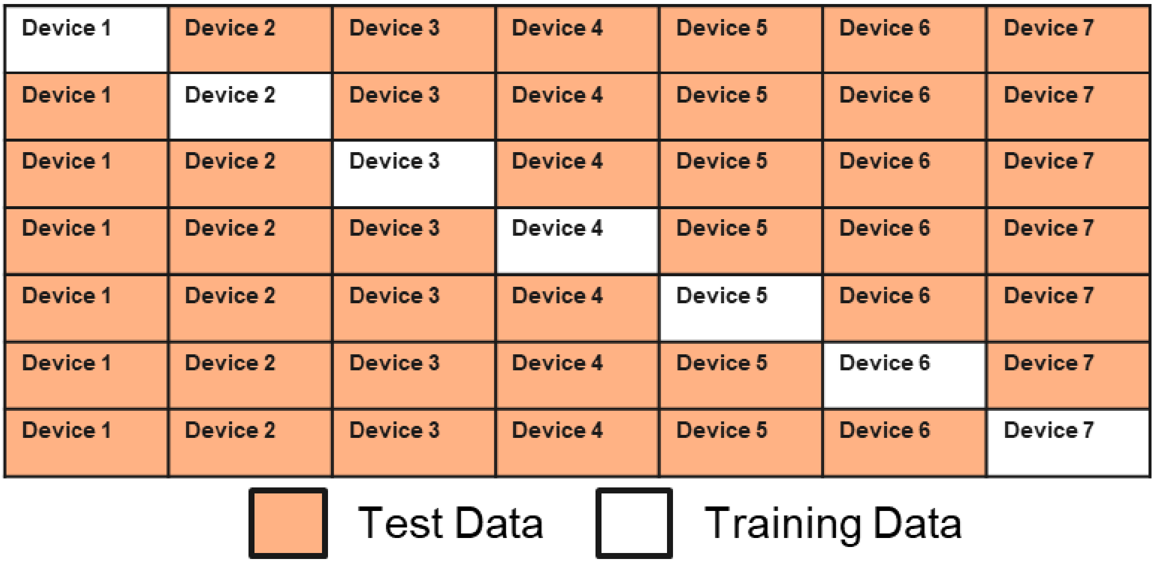
Precision is the ratio of true positives to the total number of predicted positives, while recall is the ratio of true positives to the total number of actual positives. The F1 score is a measure of the overall accuracy of a binary classification model that takes into account both precision and recall. F1 score is the harmonic mean of precision and recall and is given by:

$$\text{F1-score} = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN}, \tag{3.5}$$

The F1 score ranges between 0 and 1, with a score of 1 indicating perfect precision and recall and 0 indicating poor performance. A higher F1 score indicates a better overall performance of the classification model. In other words, the F1 score balances the tradeoff between precision and recall and is useful when the classes are imbalanced. It provides a single number that summarizes the overall performance of a binary classification model.

### *3.4.2 k-Fold Cross Validation*

Apart from selecting suitable performance assessment metrics, it is crucial to evaluate the performance of the machine learning model on various test datasets. Consequently, the $k$-fold cross-validation technique is highly recommended. In this technique, the entire dataset is initially partitioned into $k$ folds. Subsequently, one of the $k$ folds is used to train the model, while the remaining $(k-1)$ folds serve as test datasets. Lastly, the outcomes from all the considered evaluation metrics are



| | | | | | | |
|---|---|---|---|---|---|---|
| Device 1 | Device 2 | Device 3 | Device 4 | Device 5 | Device 6 | Device 7 |
| Device 1 | Device 2 | Device 3 | Device 4 | Device 5 | Device 6 | Device 7 |
| Device 1 | Device 2 | Device 3 | Device 4 | Device 5 | Device 6 | Device 7 |
| Device 1 | Device 2 | Device 3 | Device 4 | Device 5 | Device 6 | Device 7 |
| Device 1 | Device 2 | Device 3 | Device 4 | Device 5 | Device 6 | Device 7 |
| Device 1 | Device 2 | Device 3 | Device 4 | Device 5 | Device 6 | Device 7 |
| Device 1 | Device 2 | Device 3 | Device 4 | Device 5 | Device 6 | Device 7 |

🟧 Test Data □ Training Data

Figure 3.4: Cross Validation Across 7 devices (train set= 6 devices,test set = 1 device) in the Context of the Proposed Approach.

averaged to represent the overall performance of the learning classifier. The k-fold cross-validation process can be summarized as follows:

1. The dataset is divided into $k$ subsets or “folds” of roughly equal size.

2. The model is trained $k$ times, each time using a different fold as the test set and the remaining $k - 1$ folds as the training set.

3. The model’s performance is evaluated on each of the $k$ test sets, and the results are averaged to give an overall estimate.

4. The final model is trained on the entire dataset using the optimal hyperparameters selected during the k-fold cross-validation process.

k-fold cross-validation has several advantages over a single train-test split evaluation. It reduces the risk of overfitting by evaluating the model on multiple test sets. Here, a 7-fold cross-validation is adopted wherein 6 devices are used for training the model and 1 device is the hold-out data for testing Fig. 3.4.

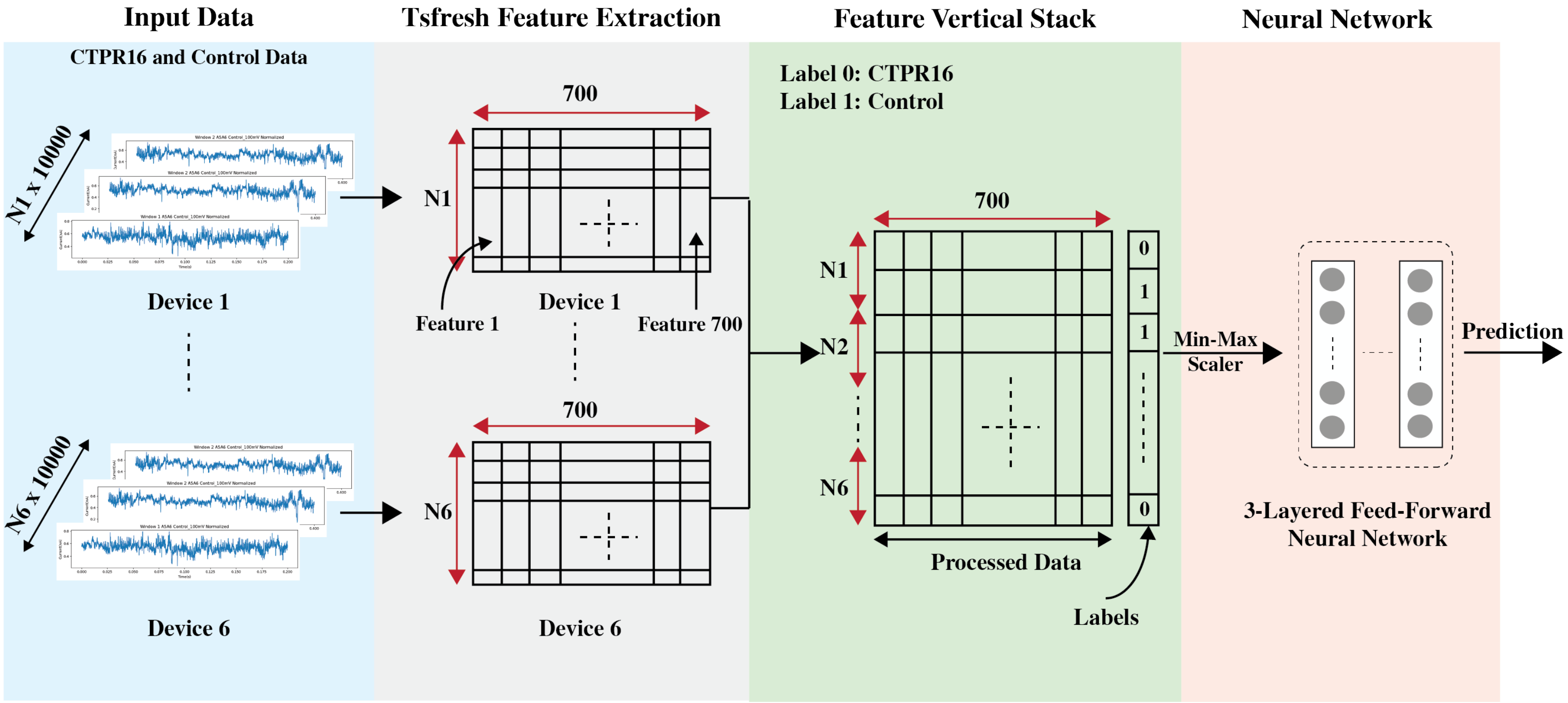


Figure 3.5: Pipeline for Feature Engineering and Deep Learning Model Training. The blue panel shows the time-series windows after pre-processing for 6 devices (training data) leaving 1 device out (test data). There are N1 number of windows for device 1, N2 for device 2 and so on. The gray panel describes 700 selected tsfresh features extracted from each time-series window and stacked into a N1 x 700 matrix for device 1, N2 x 700 matrix for device 2 and so on. The green panel shows the final combined training data as a (N1+N2+N3+N4+N5+N6) x 700 vertically stacked matrix alongwith the class labels. Label '0' denotes 'CTPR16' data and Label '1' denotes 'Control' data. The last panel shows the 3-layered feed-forward neural network used for training on 6 devices and testing on the 7th device to aid in the protein detection process.

### *3.4.3 Model Architecture*

A feed-forward neural network (FFNN) model is used to address the binary (two classes) classification problem. The FFNN model is a form of deep learning architecture that is characterized by multiple layers of interconnected neurons, enabling the model to learn complex patterns within the data. Implementation of the model required the application of TensorFlow Keras, a high-level deep learning library that builds upon TensorFlow, offering a user-friendly and efficient interface for constructing and training neural network models. The neural network architecture comprises one input layer, three hidden layers, and one output layer. The input layer accommodates 700 input features, which correspond to the selected features extracted from the dataset. Each of the three hidden layers has 1024 neurons and employs the rectified linear unit (ReLU) activation function. The ReLU activation function is widely adopted in deep learning models due to its simplicity and efficacy in addressing the vanishing gradient issue. The vanishing gradient problem can occur during the training of deep neural networks, where the gradient of the loss function (explained in Section 1.3.3 ) with respect to the network's parameters becomes very small as it propagates back through the layers of the network. This can make it challenging for the network to learn, resulting in slow convergence or poor performance. Mathematically, the ReLU activation function can be defined as:

$$f(x) = max(0, x), \tag{3.6}$$

where $x$ is the input to the neuron and $f(x)$ is the output. The derivative of the ReLU function can therefore be written as:

$$f'(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{3.7}$$

This means that the derivative of the function is always non-zero for positive input values, which can help maintain a large gradient during backpropagation. In contrast, other activation functions like the sigmoid or hyperbolic tangent functions can saturate for large or small input values, causing the gradient to become very small and leading to the vanishing gradient problem. The output layer utilizes the sigmoid activation function, which is well-suited for binary classification tasks since it maps output values to a range between 0 and 1, thereby representing the predicted probability of each class. Adam optimizer ([5]) was selected as the optimization algorithm that helps in reducing the overall loss and improving accuracy of the model. Adam (Adaptive Moment Estimation) is an optimization algorithm used in machine learning for training neural networks. It is an extension of the stochastic gradient descent (SGD) [6] optimizer that computes individual adaptive learning rates for different parameters from estimates of the first and second moments of the gradients. The Adam optimizer uses both the gradient and its momentum to update the weights during the training process. The momentum term helps to smooth out the updates and accelerate learning, while the adaptive learning rate helps to handle sparse gradients and non-stationary objectives. The learning rate of the optimizer is initialized at 0.001 and decays by a factor of 0.1 every five epochs. This adaptive learning rate enables the model to converge more rapidly and efficiently. The loss function employed during training is binary cross-entropy (explained in Equation 1.4 in Section 1.3.3), which quantifies the disparity between predicted probabilities and true labels. The model is trained over 20 epochs with a batch size of 32, facilitating iterative weight updates during the training process. In machine learning, an epoch is a complete pass through the entire training dataset in batches during the training of a model. In other words, one epoch is defined as one iteration through the entire dataset, where each sample is used precisely once for training.

### *3.4.4 Hyperparameter Optimization*

Hyperparameter optimization is a crucial aspect of developing a machine-learning model. It entails fine-tuning parameters such as the number of hidden layers, neurons, learning rate, and activation functions to optimize the model's performance for the given task. In this study, Keras Tuner, an open-source library designed explicitly for hyperparameter optimization in TensorFlow Keras models, was employed to identify the optimal hyperparameters for the FFNN model. By systematically exploring various hyperparameter combinations and assessing their performance on the validation dataset, Keras Tuner assists in pinpointing the most suitable hyperparameter configuration for the model, ultimately boosting its generalization capabilities and predictive accuracy.

## 3.5 Performance Evaluation

### *3.5.1 Accuracy and F1-Score Comparison*

This section presents the results of our experiments using a feed-forward neural network (FFNN) for classification and discuss the implications of our findings. The purpose of the study was to evaluate the performance of an FFNN classifier on a given dataset and to identify any trends or patterns that emerged from the analysis. The results indicate that the FFNN classifier performed well on the given dataset, achieving a satisfactory level of accuracy, precision, recall, and F1-score (described in Section 3.4.1) for both classes. This suggests that the chosen architecture, activation functions, and training parameters were effective in capturing the underlying patterns in the data.

| **Test Device Number** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|---|
| **1** | 0.851 | 0.855 | 0.851 | 0.853 |
| ***2**** | *0.320* | *0.610* | *0.320* | *0.232* |
| **3** | 0.672 | 0.538 | 0.672 | 0.550 |
| **4** | 0.770 | 0.760 | 0.770 | 0.750 |
| **5** | 0.791 | 0.798 | 0.791 | 0.794 |
| **6** | 0.850 | 0.860 | 0.850 | 0.850 |
| ***7**** | *0.621* | *0.496* | *0.621* | *0.551* |
| **Mean** | 0.787 | 0.762 | 0.787 | 0.759 |
| **Standard Deviation** | 0.066 | 0.118 | 0.066 | 0.111 |

Table 3.1: Performance Evaluation of the Proposed Solution. *Devices 2 and 7 when tested did not yield stable test accuracy hence those values are not included in the Mean and Standard Deviation Calculations. Stability can be attained by further optimizing the model training approach. The description of the metrics can be found in Section 3.4.1.

As can be seen from Table 3.1 that the model attained an average accuracy and F1-score (described in Section 3.4.1) of 0.787 and 0.759 with a standard deviation (SD) of 0.066 and 0.111, respectively (with an exception of device 2 and 7). An example of the the model prediction improvement with epochs for device 5 is shown in Fig. 3.7. As a reference, the initial current vs. time series data for device 5 which was used to generate features and used as a test dataset is also shown Fig. 3.6.

Possible avenues for future work include exploring different neural network architectures, such as convolutional or recurrent networks, to improve classification performance. Additionally, incorporating a more rigorous pre-processing technique, feature selection or dimensionality reduction techniques may lead to better model generalization and reduced training time. The next section investigates the dataset's features and their relationships with the target classes.

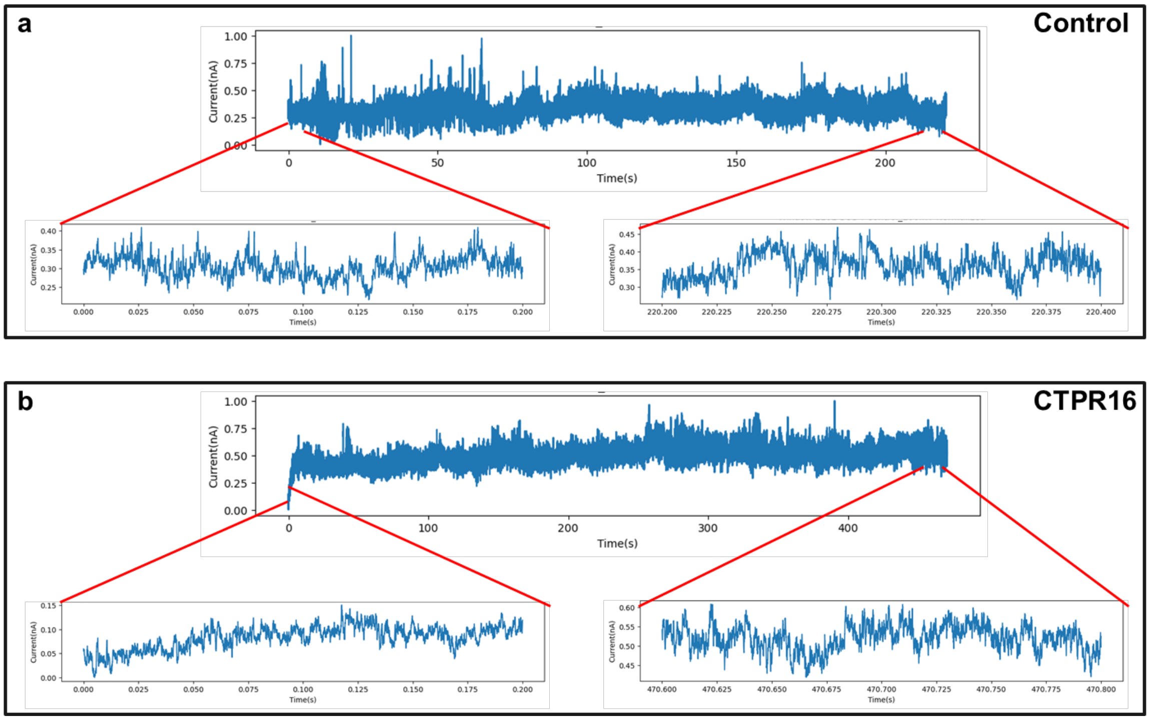


Figure 3.6: Time-series Data for the Test Device. Here, device 5 out of the 7 devices is shown as an example. The panel (a) presents the normalized current vs. time series 'Control' data for device 5 with the first and last window zoomed in. The panel (b) presents the normalized current vs. time series 'CTPR16' data for device 5 with the first and last window zoomed in. At a glance, one distinction between the two classes can be observed. It can be seen that the amplitude of current increases with time for the 'CTPR16' data whereas the amplitude levels for the 'Control' data do not change significantly with time. More complex features are extracted from these windows using 'tsfresh' and the trained model is employed on the test data features for correct prediction.

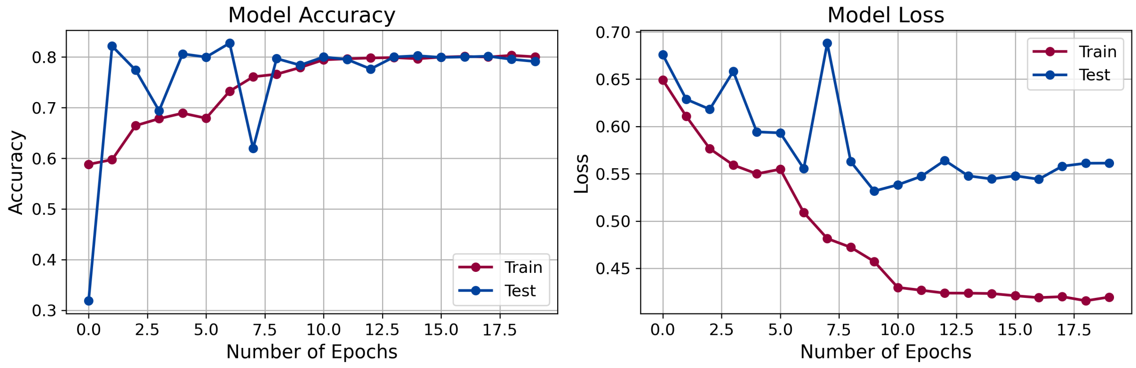


Figure 3.7: Model Accuracy and Loss vs. Epochs show a gradual improvement in the model's classification performance.

### *3.5.2 Interpreting Model Outcomes and Understanding Feature Importance with Shapley Values*

In this classification between the two experimental conditions based on the curated current vs. time dataset, Shapley values calculated through the tree-SHAP algorithm attempt to explain why an ML model reports the outputs that it does on an input. These values offer a method for fairly distributing gains and costs among various features used to predict model outcomes. Essentially, they aid in understanding how the model reaches a decision for a specific prediction. The tree-SHAP technique, developed by Lundberg et al. (2020) [7], is utilized to compute Shapley values in this study. An example of feature importance on model training when the test set is device 5 is shown in Fig. 3.8. These values are computed by altering input features by a small amount and observing how these changes relate to the final model prediction. The Shapley value for a specific feature is then computed as the average marginal contribution to the overall model score. Here, an example for test device 5 is shown wherein the training data contains current signatures from the other six devices (devices 1, 2, 3, 4, 6 and 7).

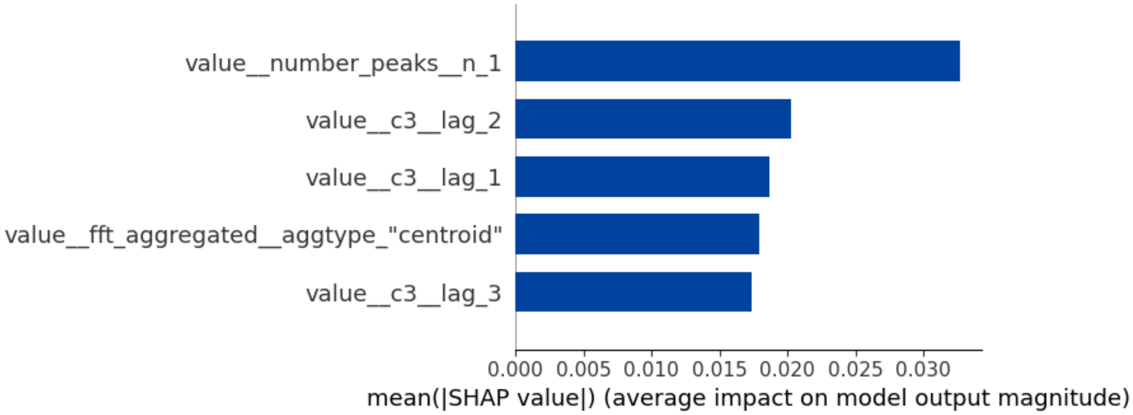


Figure 3.8: Feature Importance Ranking for the Trained Model Using Shapley Values for a Sample Device. For clarity, features are sorted by the magnitude of feature importance.

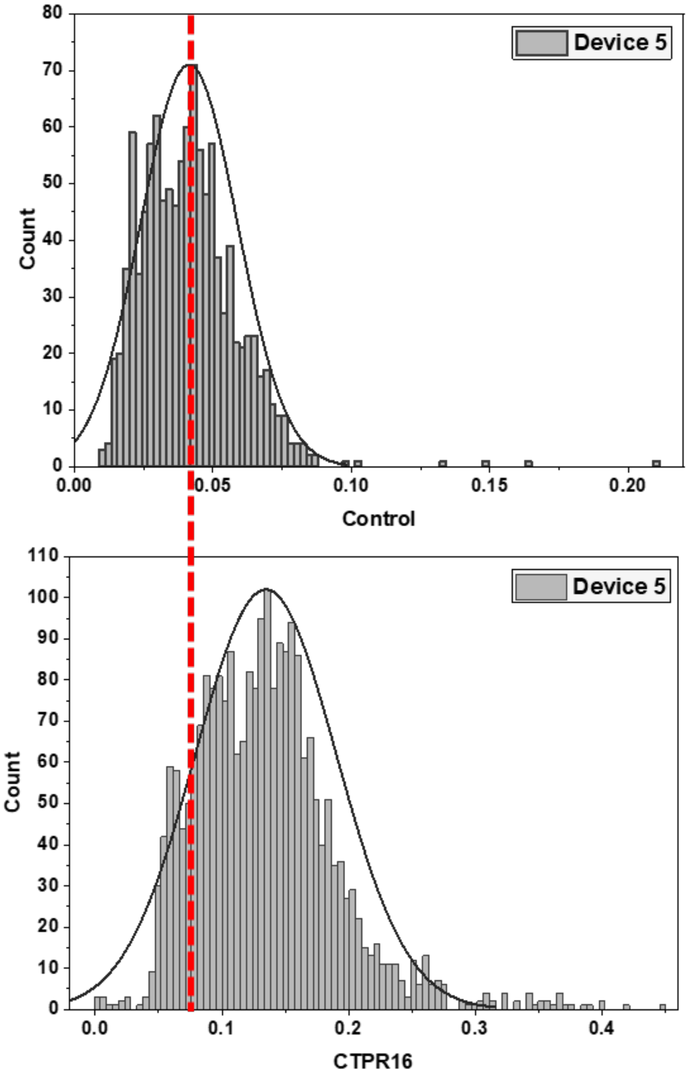


Figure 3.9: Class-wise Analysis of Important Features. Class-wise difference of feature values for the SHAPLEY-computed second most important feature (‘value c3 lag 2’ , which is a measure of nonlinearity in time series data) for the example case of test device = 5 and training data devices = 1,2,3,4,6 and 7. (Top) ‘Control’ and (Bottom) ‘CTPR16’

Focusing on one of the SHAPLEY features, the ‘c3 lag 2’ value, a clear distinction between the distributions of the ‘c3 lag 2’ value can be seen (Fig. 3.9) between the two classes of ‘Control’ and ‘CTPR16’. As a brief overview, third-order cumulant (C3) is a statistical measure that helps in analyzing non-linear time series. Non-linear time series are time series data where the relationship between past and future values is not linear and may involve higher-order interactions or dependencies. The third-order cumulant quantifies the degree of asymmetry in the distribution of a dataset. It can provide insights into the non-linear structure and dependencies in the time series data, which may be absent from linear measures such as mean and variance. The C3 measure is defined as follows:

$$C3(k,l) = \mathbb{E}\left[(x(t) - \mathbb{E}[x(t)])(x(t-k) - \mathbb{E}[x(t-k)])(x(t-l) - \mathbb{E}[x(t-l)])\right], \quad (3.8)$$

where $\mathbb{E}$ denotes the expected value, $x(t)$ is the value of the time series at time $t$; $k$ and $l$ are the lag values, and $x(t-k)$ and $x(t-l)$ are the values of the time series at times $t-k$ and $t-l$, respectively. By calculating the C3 measure for different lags ($k$ and $l$), the nonlinear dependencies in the time series can be investigated at various time scales. When the third-order cumulant (C3) is significantly different from zero, it indicates the presence of non-linearities in the data. On the other hand, if the C3 measure deviates significantly from zero for certain lags, it suggests the presence of nonlinear in the time series. Stochastic processes can be either linear or non-linear in nature. A stochastic process is a collection of random variables representing the evolution of a system over time, where the future state depends on the current state and possibly on past states. Non-linear stochastic processes might be prevalent in various real-world phenomena, such as biological systems, where the underlying dynamics are often complex and involve interactions between multiple factors. Other statistical measures shown in Fig. 3.8 included the following:

- **'value number peaks n 1':** It is defined as the number of peaks of at least support 'n' in the time series 'x'. A peak of support 'n' is defined as a subsequence of 'x' where a value occurs, which is bigger than its 'n' neighbours to the left and to the right.

- **'value fft aggregated aggtype centroid ':** It is defined as the spectral centroid (mean) of the absolute fourier transform spectrum.

These features also led to significantly different distributions for the two classes. This might be valuable in gathering further knowledge from single-molecule temporal current changes that still suffer from background noise and do not show apparent telegraph noise switching as shown in STM data. In the solid-state chip, a difference between the two kinds of dense signals might not be clearly visible at first. But, a carefully curated algorithm studying the intricacies of the time-series signal patterns shows that there is a clear effect of having molecules in the junction, and its very different from what is seen in the STM. Importantly, this work shows that there is valuable and meaningful information in what had been regarded as "failed" experiments.

## 3.6 Conclusion

In conclusion, this chapter has described using machine learning to classify between two experimental conditions, with and without CTPR16 protein wires in solution. This work emphasized on finding real data buried in the measurements that was considered as “failed” experiments. Explainable AI (Artificial Intelligence) dissects the trained model and helps find the features significantly associated with the presence of a protein bridge in the experimental setup. The work shown here is a pilot project, wherein more sophisticated methods of pre-processing and solving data imbalance can be applied to improve the model performance. Significant optimization and reproducibility can help generalize the model for future use. The applications of such a design are plenty, leaving room for customization and in-depth data exploration.

# REFERENCES

[1] Masateru Taniguchi, Shohei Minami, Chikako Ono, Rina Hamajima, Ayumi Morimura, Shigeto Hamaguchi, Yukihiro Akeda, Yuta Kanai, Takeshi Kobayashi, Wataru Kamitani, et al. Combining machine learning and nanopore construction creates an artificial intelligence nanopore for coronavirus detection. *Nature Communications*, 12(1):3726, 2021.

[2] Akihide Arima, Makusu Tsutsui, Takashi Washio, Yoshinobu Baba, and Tomoji Kawai. Solid-state nanopore platform integrated with machine learning for digital diagnosis of virus infection. *Analytical Chemistry*, 93(1):215–227, 2020.

[3] Mohamed Soliman Halawa, Rebeca P. Díaz Redondo, and Ana Fernández Vilas. Kpis-based clustering and visualization of hpc jobs: A feature reduction approach. *IEEE Access*, 9:25522–25543, 2021.

[4] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*, 307:72–77, 2018.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[6] Nikhil Ketkar and Nikhil Ketkar. Stochastic gradient descent. *Deep learning with Python: A hands-on introduction*, pages 113–132, 2017.

[7] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

Chapter 4

# DIRECT ELECTRICAL MONITORING OF THE DNA: DNA POLYMERASE INTERACTION

A part of this chapter is adapted from the manuscript titled Zhang, B.$^{\#}$; Deng, H.$^{\#}$; **Mukherjee, S.**; Song, W., Wang, X.; and Lindsay, S. Engineering an enzyme for direct electrical monitoring of activity. *ACS Nano*, **2020**, *14(2)*, 1360-1368. [1]

## 4.1 Abstract

Proteins have been shown to be electrically conductive if tethered to an electrode by means of a specific binding agent, allowing single molecules to be wired into an electrical sensing circuit. Such circuits allow enzymes to be used as sensors, detectors, and sequencing devices. We have engineered contact points into a Φ29 polymerase by introducing biotinylatable peptide sequences. The modified enzyme was bound to electrodes functionalized with streptavidin. Φ29 connected by one biotinylated contact, and a second nonspecific contact showed rapid small fluctuations in current when activated. Signals were greatly enhanced with two specific contacts. Features in the distributions of DC conductance increased by a factor 2 or more over the open to closed conformational transition of the polymerase. Polymerase activity is manifested by a rapid (millisecond) large (25% of background) current fluctuations imposed on the DC conductance.

*In this chapter, the DNA polymerase engineering was done by Dr. Hanqing Deng. The STM measurements were carried out by Dr. Bintian Zhang. The solid-state chip based experiments were carried out by Dr. Weisi Song. My contribution included the algorithm design and the statistical analysis of the experimental data.*

## 4.2 Introduction

As described in Section 1.1.2, recent studies demonstrated that several proteins could conduct electricity well when contacted by binding agents that inject charge carriers into their interiors [2, 3], despite having no redox-active centers. This conductance is electron or hole mediated and was measured under potential control in conditions that eliminated Faradaic currents [3]. The initial studies used proteins with multiple binding sites, such as antibodies or streptavidin, to bridge electrodes. However, this limits the degree to which conformational changes can be studied because the active sites are tied up as fixed electrical contact points. In the specific case of streptavidin, which has four binding sites for biotin, it is possible to use two of the sites as contacts and to study the change of conductance as additional biotin molecules bind [4]. Solving this problem for an arbitrary protein would provide a valuable tool for developing bioelectronic devices in which direct electrical measurements exploit the chemical versatility of enzymes. This study aimed to measure the current passing through a functional polymerase. This measurement motivated the construction of an enzyme with two contact points that would not interfere with the active site of a protein, mimicking a multivalent protein while retaining the biochemical activity of the original protein. Φ29 polymerase, a well-studied [5] and highly accurate [6, 7] DNA-dependent DNA polymerase already used as a single-molecule sequencing device [8], was chosen for this purpose.

Φ29 DNA polymerase is an enzyme that replicates DNA in a highly processive and

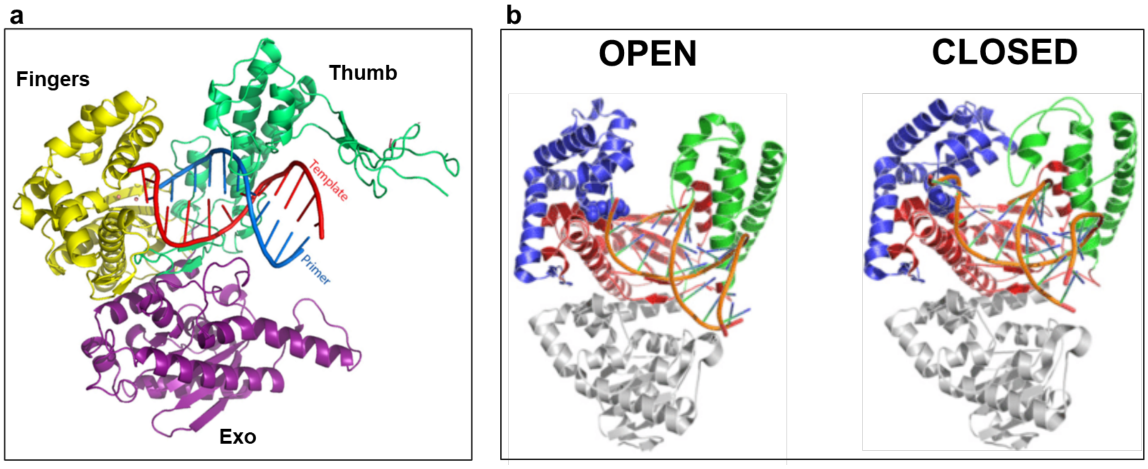


Figure 4.1: Visual Representations of the Φ29 DNA Polymerase. (a) The active domain of the polymerase resembles a human hand, with a "thumb" subdomain that holds the DNA, a "palm" subdomain that contains the catalytic site, and a moving "fingers" subdomain that closes around the complex of the DNA template once the correct complementary nucleotide triphosphate (dNTP) is bound. (b) The enzyme is normally "open" [9] and remains so after binding DNA containing a primer strand and template strand with a 5′ -overhang. (c) Once the correct dNTP is bound, the fingers close to complete the reaction, opening again only for long enough to bind the next complementary dNTP [9].

accurate manner. The enzyme undergoes conformational changes between open and closed states during its catalytic cycle, which are crucial for its function (described in Section 4.3.2.3). In the closed conformation, the Φ29 DNA polymerase is actively engaged in synthesizing new DNA. The enzyme tightly binds the template DNA strand and the incoming deoxyribonucleotide triphosphate (dNTP). The closed state is characterized by the proper alignment of the active site residues, which allows the enzyme to catalyze the formation of a phosphodiester bond between the incoming dNTP and the growing DNA strand. In the open conformation, the Φ29 DNA polymerase is disengaged from the template DNA, and the active site is not optimally aligned for catalysis. In this state, the enzyme may be involved in substrate binding, proofreading, or releasing the synthesized DNA strand. The open conformation allows for the necessary adjustments during the polymerization process, such as the translocation of the enzyme along the template DNA, the release of the pyrophosphate product, and the binding of the next incoming dNTP. Fig. 4.1 shows the visual representations of the two possible conformations of Φ29.

This chapter demonstrates a technique for inserting binding sites into Φ29 polymerase, showing that polymerase activity results in rapid conductance fluctuations using STM. Further experiments are conducted using solid-state devices to achieve two objectives. The first objective is to develop an algorithm that computes the dwell times of the polymerase in both the open and closed conformation during the processing of linear DNA templates. The second objective is to identify repeating patterns in the current vs. time data that arise from cyclic DNA templates interacting with the Φ29 polymerase. These methods highlight the advantages of utilizing bioelectronic devices for potential sequencing applications through direct electrical measurements of engineered enzymes.

## 4.3 Direct Electrical Monitoring Using STM

This section presents the methodology, experimental results and data interpretation for single molecule conductance measurements carried out using the Scanning Tunneling Microscopy.

### *4.3.1 Experimental Design*

#### *4.3.1.1 Recombinant Φ29 DNA Polymerase Constructs with Inserted Avitag*

The starting enzyme utilized was a Φ29 DNA polymerase, which was rendered exonuclease-deficient by introducing D12A and D66A mutations. The Avitag DNA sequence was inserted into a pET15b plasmid containing the mutant polymerase gene using a Q5 site-directed mutagenesis kit (NEB). The corresponding inserted peptide sequence is depicted in blue, while the flanking linker sequences are shown in yellow in Fig. 4.2 a. The $\epsilon$-amine of the central lysine (K), indicated by the red arrow in Fig. 4.2 a, was biotinylated using the BirA enzyme [10]. Three generations of the modified enzyme were examined. The first generation (Gen I) was biotinylated exclusively at the N-terminus of the protein. The second generation (Gen II) included an additional Avitag approximately 5 nm away from the N-terminus, positioned between E279 and D280. This secondary site is situated in the deactivated exonuclease domain and was selected due to its stable position relative to the N-terminus during the open-to-closed transition [11, 12]. The third generation (Gen III) incorporated an extra flexible linking sequence (GNSTNGTSNGSS) adjacent to the N-terminal Avitag, allowing for increased flexibility in contact geometry. Biotinylation was confirmed through SDS-PAGE gel analysis of both the free and streptavidin-bound polymerases. Fig. 4.2 b illustrates the molecular weight increase resulting from streptavidin binding to the Gen III polymerase (with higher molecular weight features likely representing polymer aggregates alternating between polymerase and streptavidin). An activity test of all versions of Φ29 DNA polymerase was conducted using a rolling circle replication reaction. The reaction mixture consisted of 1.25 pmols a RCR template and primer complex (mentioned in Section 4.3.1.3, 500 $\mu$M dNTP, and 4 pmols Φ29 DNA polymerase in 1X reaction buffer containing 50 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 10 mM $(NH_4)_2SO_4$, and 4 mM DTT. This mixture was incubated at 30°C for 1 hour. GelRed (Biotium) staining was used to visualize the product on a 0.8% agarose gel. The polymerization products of the modified Φ29 are displayed in lane 1 of Fig. 4.2 c. When the Φ29 is complexed with streptavidin, the production of polymers remains virtually unchanged (lane 4, Fig. 4.2 c). The assays were repeated in the buffer used for STM measurements (1 mM phosphate buffer, pH=7.4, 4 mM TCEP, 10 mM $MgCl_2$ with 1 mM dNTPs and 1 $\mu$M template) to confirm that activity was maintained.

#### *4.3.1.2 Functionalization of STM Substrates and STM Probes*

As discussed in Chapter 2.2.1 and shown in Fig. 2.1, palladium substrates were fabricated by evaporating a 200 nm palladium film onto a silicon wafer using an electron-beam evaporator (Lesker PVD 75) and a 10 nm titanium adhesion layer. Prior to functionalization, the substrates were treated with a hydrogen flame and subsequently immersed in solutions of thiolated streptavidin (ProteinMods) or thiolated biotin overnight. Thiolated biotin was synthesized as described elsewhere [3] and dissolved in freshly degassed pure ethanol at a final concentration of 50 $\mu$M. A 1 $\mu$M thiolated streptavidin solution in 1 mM PB buffer was employed for substrate functionalization. All buffers and solutions were prepared using Milli-Q water with a resistivity of 18.2 M$\Omega cm$ . For all experiments, the 1 mM PB buffer (pH 7.4) was degassed with argon to prevent oxygen interference. The polymerization buffer

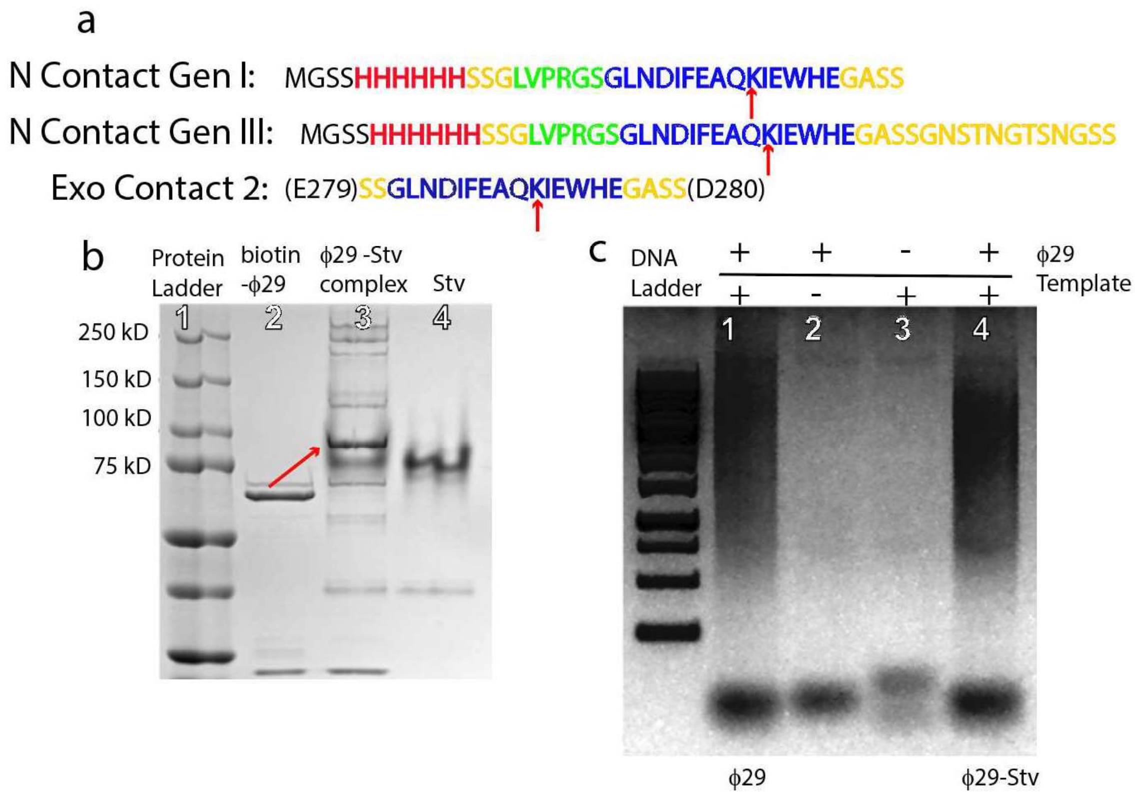

a
N Contact Gen I: MGSSHHHHHHSSGLVPRGSGLNDIFEAQKIEWHEGASS
N Contact Gen III: MGSSHHHHHHSSGLVPRGSGLNDIFEAQKIEWHEGASSGNSTNGTSNGSS
Exo Contact 2: (E279)SSGLNDIFEAQKIEWHEGASS(D280)
b
Protein Ladder
biotin -ϕ29
ϕ29 -Stv complex
Stv
1
2
3
4
250 kD
150 kD
100 kD
75 kD
c
DNA Ladder
+
+
-
+
ϕ29
+
-
+
+
Template
1
2
3
4
ϕ29
ϕ29-Stv

Figure 4.2: Protein Sequences, Synthesis and Gel Analysis. (a) Avitag sequence (blue) inserted at the N terminus (N Contact Gen 1), (b) The formation of a complex with streptavidin was verified by a protein gel which shows how the biotinylated Φ29 (lane 2) forms a complex with streptavidin molecule (lane 3). Lane 4 is streptavidin alone and, (c) Activity of modified polymerase complexed with streptavidin.

consisted of 1 mM phosphate buffer (pH 7.4), 4 mM tris(2-carboxyethyl)-phosphine) (TCEP), 10 mM $MgCl_2$, 1 mM dNTPs, and 1 $\mu$M template (activity in this buffer was confirmed through a rolling circle amplification assay). STM probes were etched from a 0.25 mm Pd wire (California Fine Wires) using an AC electrochemical etching method. To circumvent current leakage, probes were insulated with high-density polyethylene following the procedure previously described for gold probes [13]. Each probe was tested via STM in 1 mM phosphate buffer (pH 7.4) at a +0.5 V bias to ensure the leakage current was below 1 pA. Probes were functionalized with streptavidin using either a thiolated biotin (SH-biotin) solution or thiolated streptavidin solution for 4 hours or overnight, rinsed with water, dried with nitrogen gas, and used immediately.

a

5'CCCCCCCCCC AACTGGCCG TCGTTTTACA TATGTAAAC GACGGCCAGT T3'

```
┌ACATTTTGCTGCCGGTCAACCCCCCCCCC 5′
T    |||||||||||||||
└ATGTAAACGACGGCCAGTT
```

b

5'AAAAAAAAAA AACTGGCCG TCGTTTTACA TATGTAAAC GACGGCCAGT T3'

```
┌ACATTTTGCTGCCGGTCAAAAAAAAAAAA 5′
T    |||||||||||||||
└ATGTAAACGACGGCCAGTT
```

c

5'CCCCCCCCCC AAAAAAAAAA AACTGGCCG TCGTTTTACA TATGTAAAC GACGGCCAGT T3'

```
┌ACATTTTGCTGCCGGTCAAAAAAAAAAAACCCCCCCCCC 5′
T    |||||||||||||||
└ATGTAAACGACGGCCAGTT
```

d

5'ATC ATC ATC ATC ATC AACTGGCCG TCGTTTTACA TATGTAAAC GACGGCCAGT T3'

```
┌ACATTTTGCTGCCGGTCAACTACTACTACTACTA 5′
T    |||||||||||||||
└ATGTAAACGACGGCCAGTT
```

Figure 4.3: Sequences of the DNA Templates Used and Their Folded Structures.

#### *4.3.1.3 DNA Templates*

A collection of four different single-stranded template with a 15 basepair hairpin primer was used for the STM conductance measurements. The sequences of the DNA templates and their folded structures are presented in Fig. 4.3.

#### *4.3.1.4 Conductance Measurements*

STM measurements were carried out using a PicoSPM scanning probe microscope (Agilent Technologies) combined with a DAQ card (PCI-6821 or PCIE-7842R, National Instruments) for data acquisition. The Teflon liquid cell was cleaned with Piranha solution and subsequently sonicated in Milli-Q water to eliminate residues (note that Piranha solution is highly corrosive and must be handled with extreme caution). To more effectively control the surface potential, an Ag/AgCl reference electrode with a 10 mM KCl salt bridge was connected to the substrate, keeping both electrodes in the double-layer region of potential where no Faradaic currents flow [3]. Protein conformations can be affected by the double-layer electric field [14], but measurements indicated minimal effects from minor changes in surface potential, as typically employed in this study [3]. Initially, the probe was engaged at a 4 pA setpoint current with a bias of -0.2 V and allowed to stabilize for 2 hours prior to measurement. Detailed descriptions of IV sweep and current vs. time (I(t)) measurements can be found in reference [3]. For STM IV sweep measurements, the servo system was first deactivated, and the probe was retracted by ΔZ nm at a speed of 1 nm/s. The probe was then suspended at that height for 1 minute, during which a custom LabVIEW program monitored current changes. When the current surpassed a 40 pA threshold, a binding event was considered to have taken place, and IV sweeps were initiated from -0.2 V to +0.2 V and back at a sweep rate of 1 V/s, followed by a 0.2 s resting period. The current was subsequently rechecked, and if it remained above twice the noise level (6 pA), IV curves were continuously recorded until the bound protein molecule detached. After one minute of measurement, the servo system was re - engaged, and the entire process was repeated. For each measurement, at least 1000 IV curves were collected, and curves with overlapping up sweeps and down sweeps (80% of the total) were chosen to construct the conductance distribution histogram.

Current vs. time (I(t)) traces were recorded using another LabVIEW program following a similar procedure, except that the bias was held constant during the probe-holding process. The analog-to-digital sampling rate was set at 50 kHz.

Conductance measurement procedures for all analytes were identical, but with varying efficiencies due to differences in binding affinity and functionalization efficiency. The potential relative to the reference electrode was established using a battery-powered voltage source connected between the substrate and the reference electrode.

### *4.3.2 Results and Discussion*

#### *4.3.2.1 Engineering Contacts*

The design criteria for the engineered enzyme included: (1) ensuring contact points are distant from the active site of the enzyme to prevent constraining its functional motions; (2) maintaining a constant relative spacing of contact points during the open-to-closed conformational transition; (3) maximizing the spacing between contact points; and (4) preserving the isoelectric point (pI) of the enzyme, such that the inserted sequences do not alter it. To satisfy criterion (2), atomic sites with identical crystal - structure coordinates in both conformations related to enzymatic activity were chosen [9]. It is essential to do this because these points are constrained by fixed contacts. This constraint would affect polymerase activity if contact points that moved over the conformational transition were chosen. A larger spacing (criterion 3) enables better separation of contact points, thus facilitating junction construction. Ensuring that the pI of the modified protein remains close to the wild type (criterion 4) prevents alterations in protein folding due to changes in the protein surface charge distribution. Insertion of sequences that significantly altered the pI rendered the polymerase inactive. Polymerase activity measurements served to assess the viability of the engineered polymerases.

In the first iteration, the Avitag peptide sequence was inserted near the N-terminus of the polymerase (Gen I). The lysine in this sequence is biotinylated using the BirA enzyme [10, 15], allowing for robust and specific binding to streptavidin. Biotin-bound streptavidin functions as an excellent molecular wire [3] and prevents Φ29 from coming in close proximity with the metal electrodes, as it has seven surface cysteines that can bind directly to a metal electrode and denature the polymerase. A second version (Gen II) incorporated a second Avitag, approximately 5 nm away

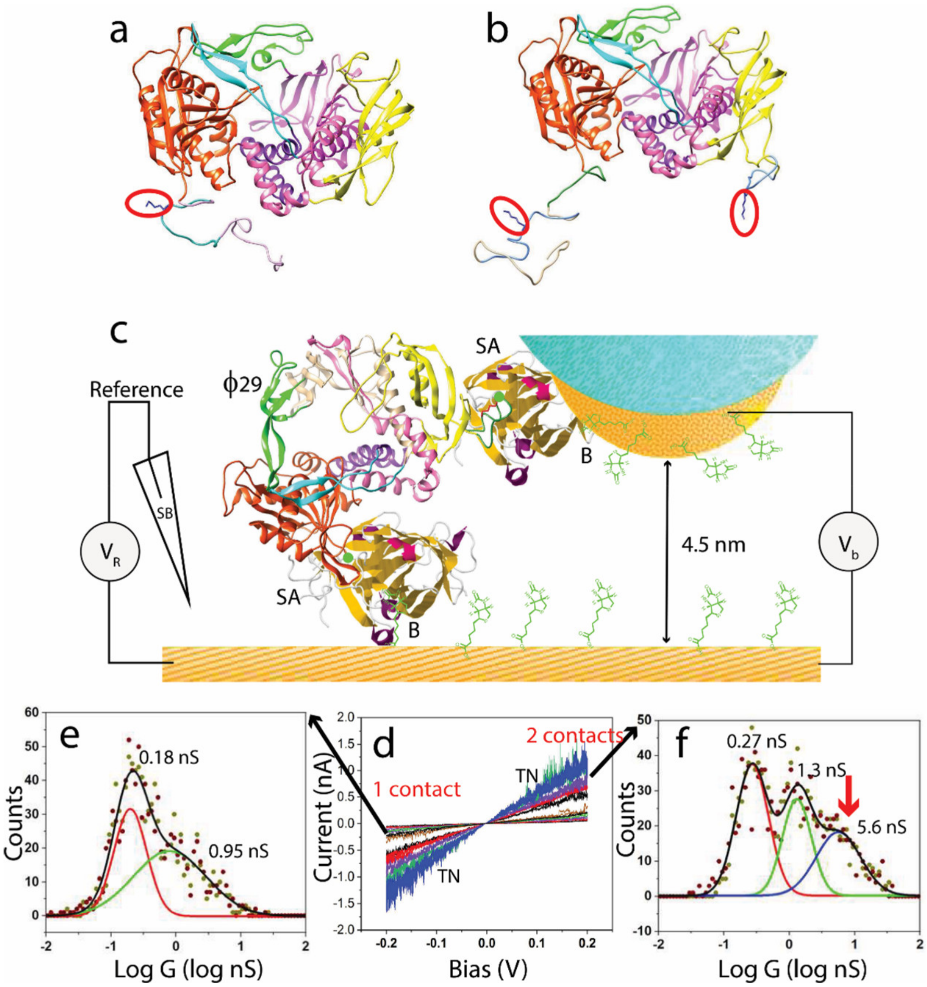


Figure 4.4: Conductance of polymerases with one and two biotinylated contact points. (a) Φ29 polymerase with a single Avitag at the N terminus (Gen I). Biotinylatable lysine is labeled by the red outline. (b) Φ29 polymerase with a second Avitag inserted between E279 and D280 and a flexible loop at N terminus (Gen III). (c) STM probe is held ∼4.5 nm above a conducting substrate, immersed in electrolyte and under potential control via a salt bridge ("SB") to an Ag/AgCl reference. Electrodes functionalized with thiolated biotin ("B") capture streptavidin molecules ("SA") which trap a biotinylated polymerase ("Φ29"). (d) Typical current - voltage curves (trace and retrace are superimposed). Conductances for individual molecules are obtained from the slopes of these traces. Telegraph Noise or "TN" indicates the region of contact field induced fluctuations. A doubly biotinylated polymerase has a new high conductance feature at ∼6 nS in the conductance distribution (red arrow in f) not present in the singly biotinylated molecule (e). The largest uncertainty in the fitted peaks is ± 0.05 in log G, corresponding to about ± 0.12G in G. Peak widths and fitting errors for each peak shown here are listed in Table 4.1.

from the first, at a site in the inactivated exonuclease domain. A third version (Gen III) contained both contacts and a flexible linker adjacent to the N-terminal Avitag. The visual representations of generations I and III are displayed in Fig. 4.4 a,b. Peptide sequences, protein gel formation of the complex with streptavidin, and the activity of the complex are described in the SI of this chapter's associated publication [1]. The same assay verified polymerase activity in the buffer employed for STM measurements.

#### *4.3.2.2 STM Conductance Measurements*

Measurements were made using an electrochemical scanning tunneling microscope (Pico STM, Agilent) with insulated palladium (Pd) probes [13] and a Pd substrate, both held under potential control using a salt-bridged reference electrode (Fig. 4.4 c). Electrodes were modified with streptavidin using either a thiolated biotin (SH-biotin) or thiolated streptavidin and then incubated with a solution of the biotinylated polymerase (Section 4.3.1). Measurements were made in a reaction buffer containing $MgCl_2$ and tris(2-carboxyethyl)-phosphine) (TCEP) to prevent polymerase oxidation. Nucleotide triphosphates were added to activate the polymerases. Current-voltage (IV) characteristics were measured using a fixed Z gap (no servo control) which remained constant to within about 0.1 nm over ∼1 min, as determined by tunnel - current measurements. Drift in the X - Y plane cannot be measured accurately, but the contact point with a target molecule changes over time. The bias was swept between -0.2 and +0.2 V and back again at a rate of 1 V/s. After 1 min, the gap was returned to the set - point value, and the cycle was repeated to obtain further IV sweeps. Eighty percent of these sweeps were linear and reproduced exactly on reversing the sweep direction (Fig. 4.4 d). The gradients of these sweeps were used to compile conductance distributions that reflect different types of contacts (Fig. 4.4

e,f).

In contrast to many electrochemical conduction processes, the conductance does not depend on the scan rate because it is electronic (Section 4.3.1). This was verified for the polymerase with repeated scans at different scan speeds. At high speeds (10 V/s), there was a significant capacitative current (i.e., hysteresis), but the slopes of the IV curves were unchanged. The capacitative contribution was insignificant at a scan rate of 1 V/s, which is why this rate was chosen.

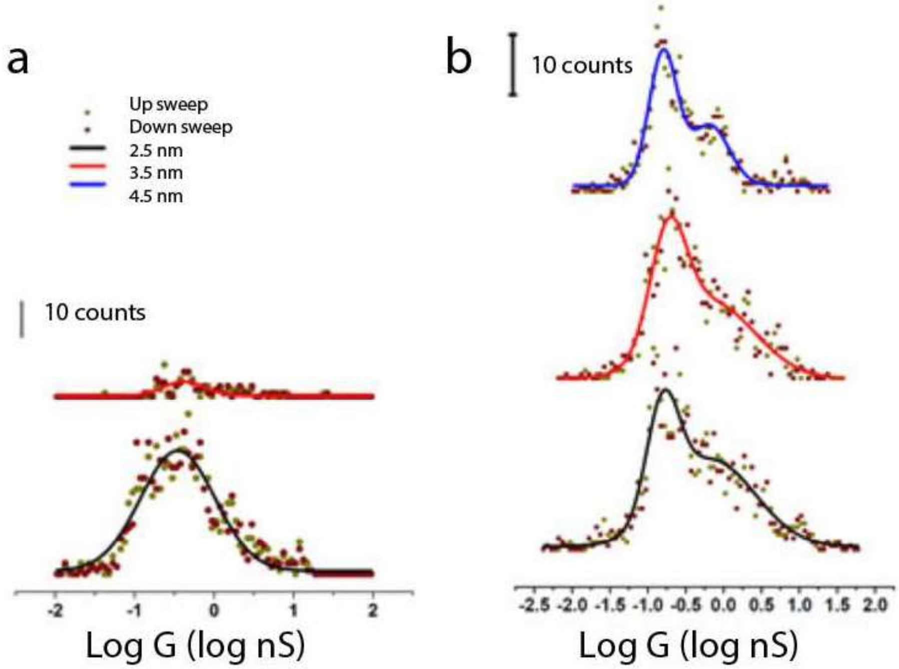


Figure 4.5: Conductance Distributions as a Function of Gap Size for (a) Streptavidin Functionalized Electrodes and (b) After the Introduction of Gen I Monobiotinylated Φ29.

The molecular junctions (Fig. 4.4 c) were assembled by coating the electrodes with streptavidin, using thiolated streptavidin or wild-type streptavidin in combination with biotinylated electrodes (Section 4.3.1). Conductance through the streptavidin alone is only observed when the gap is less than 3.5 nm. When biotinylated Φ29 is added to the liquid cell, high conductance is observed out to a 4.5 nm gap (Fig. 4.5).

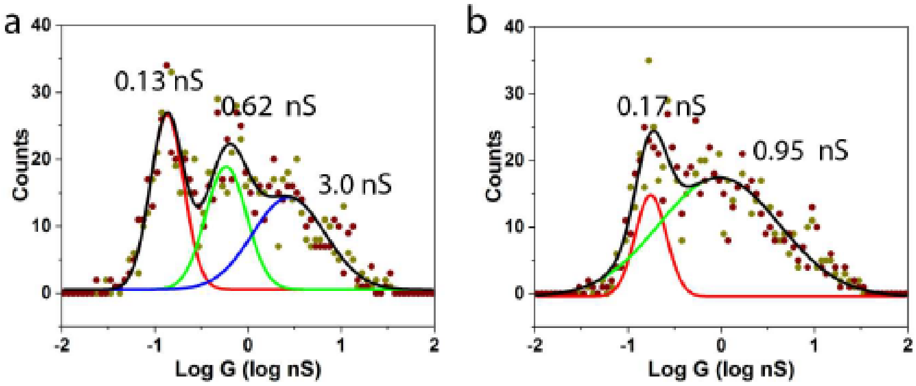


Figure 4.6: Conductance Distribution for the Gen II Bisbiotinylated Polymerase. (a) Conductance distribution for the Gen II bisbiotinylated polymerase contacted via thiolated streptavidin to the electrodes (Gap = 4.5 nm). Note that the highest conductance peak is smaller than that observed for wild-type streptavidin connected to the electrodes via thio-biotin (Fig. 4.4 f), consistent with observations of the conductance of streptavidin alone [3]. (b) Conductance distribution for the Gen II polymerase attached to a thiolated substrate and contacted with a bare probe, indicating that the two smallest peaks in the conductance distribution arise from interactions between surface thiols on the Φ29 and the bare metal.

Signals were obtained up to a 6.5 nm gap but with decreased frequency. Since the gap between contact points on the polymerase is 5.5 nm, this observation suggests that a vertical alignment of the streptavidin-polymerase-streptavidin complex is unlikely. Unexpectedly, the monobiotinylated Gen I polymerase gave two conductance peaks (Fig. 4.4 e). The first peak (at ∼ 0.2 nS) is characteristic of one specific, and one nonspecific contact [3]. The additional peak may reflect interactions between surface cysteines on the polymerase and the electrodes (Fig. 4.6 b). The bis-biotinylated Gen III displays yet another high conductance peak in addition to the two observed in the monobiotinylated polymerase (Fig. 4.4 f). This new high conductance peak has a value (∼ 5.6 nS) that is characteristic of a bridge formed by the binding of specific ligands [3]. The width of the peaks reflects the frequency with which various types of contact are made. This is illustrated by the fact that distributions of conductance obtained from monitoring current at a fixed voltage as the STM gap is left to drift in the XY plane (i.e., at constant height) reproduce the distributions obtained from repeated IV measurements [3]. The width of the lowest conduction peak is little changed between the case of one contact (Fig. 4.4 e) and two contacts (Fig. 4.4 f). However, it is narrowed considerably in the second peak for the case of two contacts, presumably because the binding of the second streptavidin shields some of the surface cysteines from the metal electrode surface. It reduces the number of ways that surface cysteines can interact with the metal electrodes (peak widths are listed in Table 4.1). Interestingly, connection via biotinylated electrodes gave rise to higher conductance than direct attachment via thiolation of surface residues (Fig. 4.6 a), a phenomenon that was observed previously for conduction through streptavidin alone [3]. It can be speculated that the binding of biotin into a deep pocket in the streptavidin provides

| **Figure/Experiment** | **Peak Number** | **Peak (Log(G))** | **Half Width (Log(G))** |
|---|---|---|---|
| **Fig. 4.4 e** | 1 | -0.74±0.01 | 0.51±0.05 |
| Single Contact | 2 | -0.02±0.03 | 1.13 ±0.17 |
| **Fig. 4.4 f** | 1 | -0.56±0.01 | 0.50±0.01 |
| Two contacts | 2 | 0.12±0.03 | 0.44 ±0.03 |
| | 3 | 0.75±0.04 | 0.68 ± 0.06 |
| **Fig. 4.7 a** | 1 | -0.75±0.02 | 0.50±0.04 |
| Pol + Template | 2 | -0.05±0.02 | 0.44 ±0.05 |
| | 3 | 0.82±0.03 | 0.58 ± 0.08 |
| **Fig. 4.7 b** | 1 | -0.62±0.02 | 0.59±0.04 |
| Pol + Template + dNTP | 2 | 0.32±0.03 | 0.61 ±0.1 |
| | 3 | 1.10±0.07 | 0.54 ± 0.11 |
| **Fig. 4.7 c** | 1 | -0.48±0.01 | 0.42±0.02 |
| Pol + Template + NHdNTP | 2 | 0.46±0.03 | 0.77±0.08 |
| | 3 | 1.18±0.02 | 0.17 ± 0.04 |

Table 4.1: Fits to the Distributions of Log (G). Peak positions and widths with the uncertainties in the fits. Units for G are nS.

a better injection of carriers into the hydrophobic interior of the protein [2].

In summary, the highest conduction peak was observed due to the conductance through the polymerase molecules. The next section describes the conformational dependence of the electronic conductance.

#### *4.3.2.3 Conformational Dependence of Polymerase Conductance*

The active domain of the polymerase resembles a human hand, with a "thumb" subdomain that holds the DNA, a "palm" subdomain that contains the catalytic site, and the moving "fingers" subdomain that closes around the complex of DNA template once the correct complementary nucleotide triphosphate (dNTP) is bound (visual representations in Fig. 4.1). The enzyme is normally "open" [9] and remains so after binding DNA containing a primer strand and template strand with a $5'$ - overhang (Fig. 4.1). Once the correct dNTP is bound, the fingers close to complete the reaction, opening again (Fig. 4.1) only for long enough to bind the next complementary dNTP [9]. This transient opening can be suppressed by using nonhydrolyzable dNTPs (NH-dNTPs) in which a carbon replaces an oxygen in the triphosphate [16]. The measurements of the conductance distributions were repeated: (a) with a saturating (1 $\mu$M) concentration [17] of a single-stranded template with a 15 basepair hairpin primer (Fig. 4.3 c), (b) with the template-bound polymerase in the presence of a saturating concentration [17] (1 mM) of dNTPs, and (c) in the presence of a saturating concentration [17] (1 mM) of NHdNTPs. The dissociation constant for $\Phi$29-template interactions is 80 nM [18]. The Michaelis-Menten constant for dNTP binding to template-bound polymerase is between 5 and 30 $\mu$M [17, 18]. All measurements were made in the presence of 10 mM $MgCl_2$. The corresponding conductance distributions are shown in Fig. 4.7 a–c. The distribution in the presence of the bound template ( Fig. 4.7 a) is almost identical to the distribution in the absence of the template (Fig. 4.7 f, uncertainties in these fits are discussed in the caption and listed in Table 4.1). On addition of dNTPs there are significant shifts in the conductance peaks (Fig. 4.7 b, 1.3×, 2.3×, and 2× for peaks 1, 2 and 3, respectively). Locking the polymerase in the closed form changes the peak positions a little more (Fig. 4.7 c, 1.8×, 3.2×, and 2.3×) with a notable sharpening of the third peak (Table 4.1 presents a list of fitted peak widths). This phenomenon can be explained as follows. The "unlocked" but dNTP-bound polymerase (Fig. 4.7 b) may be fluctuating rapidly between subconformations, broadening the high conductance peak, which then narrows substantially as the closed conformation is locked in with non-hydrolyzable dNTPs

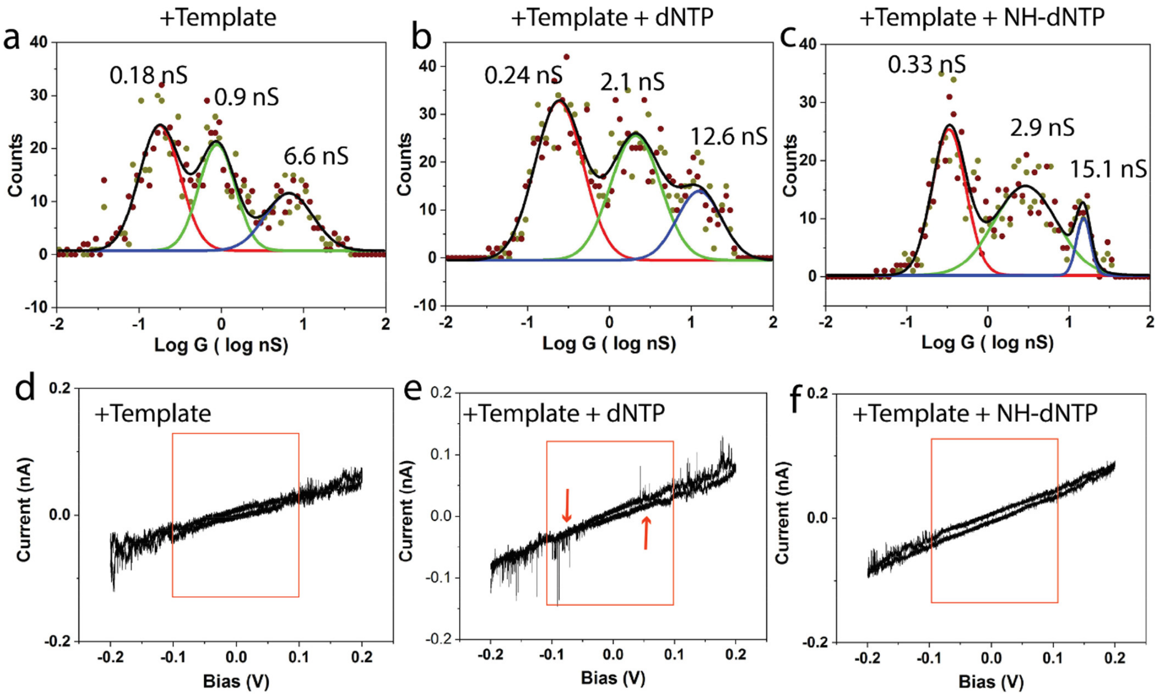


Figure 4.7: Open to closed Transition Changes Polymerase Conductance. (a) Distribution in the absence of dNTPs but with bound template. The polymerase is largely open. (b) With dNTPs added (mostly closed) the distribution changes dramatically. (c) Conductance distribution for a polymerase locked in the closed conformation with nonhydrolyzable dNTPs. Peak widths and fitting errors for each peak shown here are listed in Table 4.1. (d-f) In the inactive state (-dNTP, d, or + NH-dNTP, f) the IV curves are noise free in the bias range below ± 100 mV (red box). The active polymerase (e) shows noise spikes on the IV curve in this otherwise quiet region (red arrows). All measurements were made with 10 mM $MgCl_2$.

(Fig. 4.7 c). The observations clearly indicated that the open to closed transition of the polymerase is accompanied by large changes in conductance.

These measurements were taken in the presence of $Mg^{2+}$, so the polymerase is catalytically active in the presence of both template DNA and dNTPs. This is marked by additional noise, as shown in samples of the IV curves for the three cases in Fig. 4.7 d-f. The inactive polymerase (Fig. 4.7 d,f) does not display large noise spikes in the bias region below ±100 mV, as reported for other proteins [3]. (Above 100 mV, the electric field at the contact points induces telegraph noise; the current at a fixed bias switches between two distinct levels [3, 19].) However, when the polymerase is active (Fig. 4.7 (e)), large spikes are also observed in the bias region below ±100 mV.

#### *4.3.2.4 Noise Measurements*

These findings suggest that rapid polymerase activity can be monitored by measuring the current as a function of time ($I(t)$) at a bias below 100 mV in the presence of template-bound Φ29, $Mg^{2+}$, and dNTPs. A gap of 2.5 nm was maintained using servo control, followed by opening the servo, increasing the gap to 6 nm, and then lowering the tip to 4.5 nm to record the current for 60 s at a bias of 50 mV. Typically, no current was detected for the initial 10-20 s, after which contact formed and an $I(t)$ curve was obtained. Contacts were formed with molecules in more than 50% of these "fishing" attempts. The currents jumped suddenly upon contact with the molecule, but then changed significantly as the contact point drifted.

A representative current-time trace is depicted in Fig. 4.8 (a). This variation in current versus time was demonstrated to be a result of the drift in the contact point; the distribution of currents taken in an $I(t)$ curve replicates the distribution measured by taking IV curves from numerous different contact points [3].Telegraph noise (TN) is clearly observable in $I(t)$ traces acquired with activated polymerases.

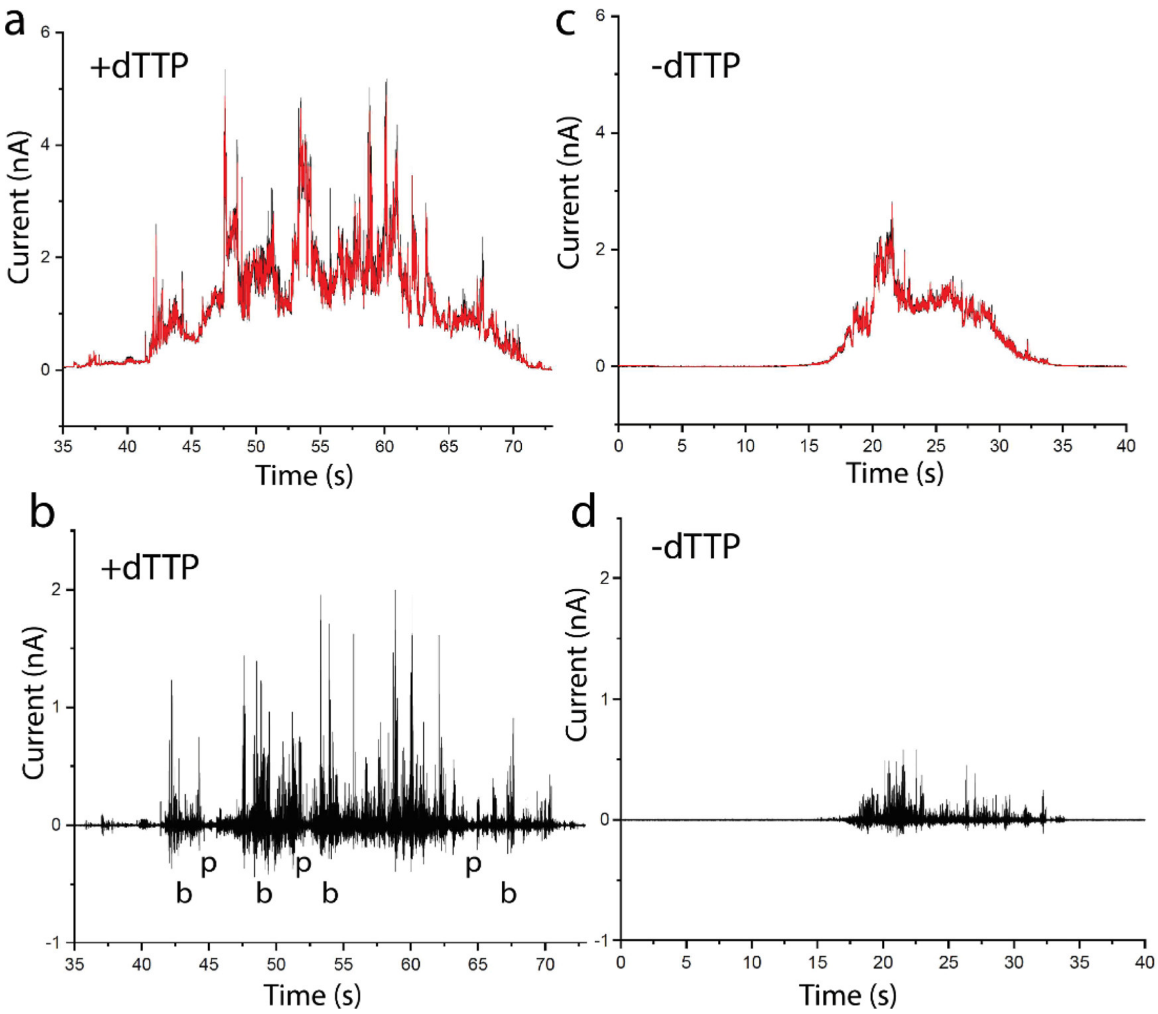


Figure 4.8: Separating signal from background conductance changes. (a) Current through the polymerase changes markedly over time as the STM probe drifts (mainly obscured black curve). This drifting baseline current is fitted with an asymmetric least-squares (ALS) procedure to yield a smoothed background current (red curve superimposed on black curve). (b) Subtraction of the fitted background shows the rapid changes in current that occur in an activated polymerase. Typically, the dynamic signals occur in bursts (b) interspersed with pauses (p). (c) Shows the current recorded vs time for this $dA_{10}$ template in the absence of dTTP. The ALS fit is shown in red. (d) The baseline subtracted signal shows that there is also noise in the inactive system though smaller in amplitude than that for the activated polymerase (these data sets (a, c) were chosen to have about the same DC conductance).

Fig. 4.9 illustrates examples for the Gen I, single contact polymerase. (The TN is aperiodic and thus does not contribute well-defined features to a power spectrum of the signal.) Nevertheless, these data also demonstrate the challenge of quantifying the noise amidst a rapidly varying background current.

To eliminate the variable background, an asymmetric least-squares (ALS) fit [20] was used. The ALS accurately follows the background without distorting the noise signals (Fig. 4.10). The $I(t)$ trace shown in Fig. 4.8 a was acquired with a $dA_{10}$ template (Fig. 4.3 (b)) in the presence of dTTP. The raw data are displayed as black points, mostly obscured by the ALS fit (red). The subtracted signal, corresponding to the fluctuations, is presented in Fig. 4.8 b. When the same procedure is applied to a trace taken at approximately the same current in the absence of dTTP (Fig. 4.8 c,d), it is evident that noise is also present in the -dTTP control. However, the noise exhibits a much smaller amplitude. A detailed examination of the signals uncovers two distinct levels of telegraph noise, as illustrated in Fig. 4.11 a and labeled “SF” (small fluctuations) and “LF” (large fluctuations). Noise-amplitude distributions for the traces in Fig. 4.8 b,d are depicted in Fig. 4.11 b. The SF appear in all measurements, whereas the LF are only present when the polymerase is active. Quantifying this

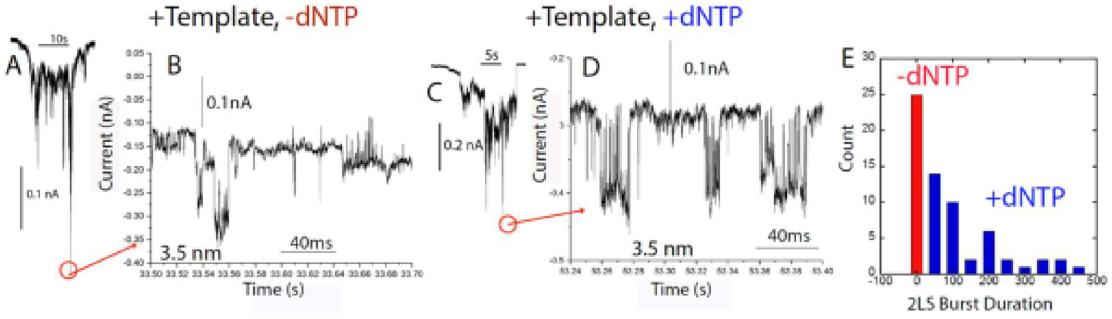


Figure 4.9: Noise Signals Obtained with the Gen I Monobiotinylated Polymerase. (A) Current vs. time with bound template but no dNTPs. The highest current region is expanded in (B). There are many jumps in the baseline but no obvious two-level telegraph noise. (C) Current-time recording with dNTPs added. Noise spikes are more evident and, when an expanded trace is plotted (D) obvious bursts of two-level random telegraph noise are present. (E) is a histogram of burst duration, clearly longer for the +dNTP case. Note: in these plots current is increasing in the downward direction.

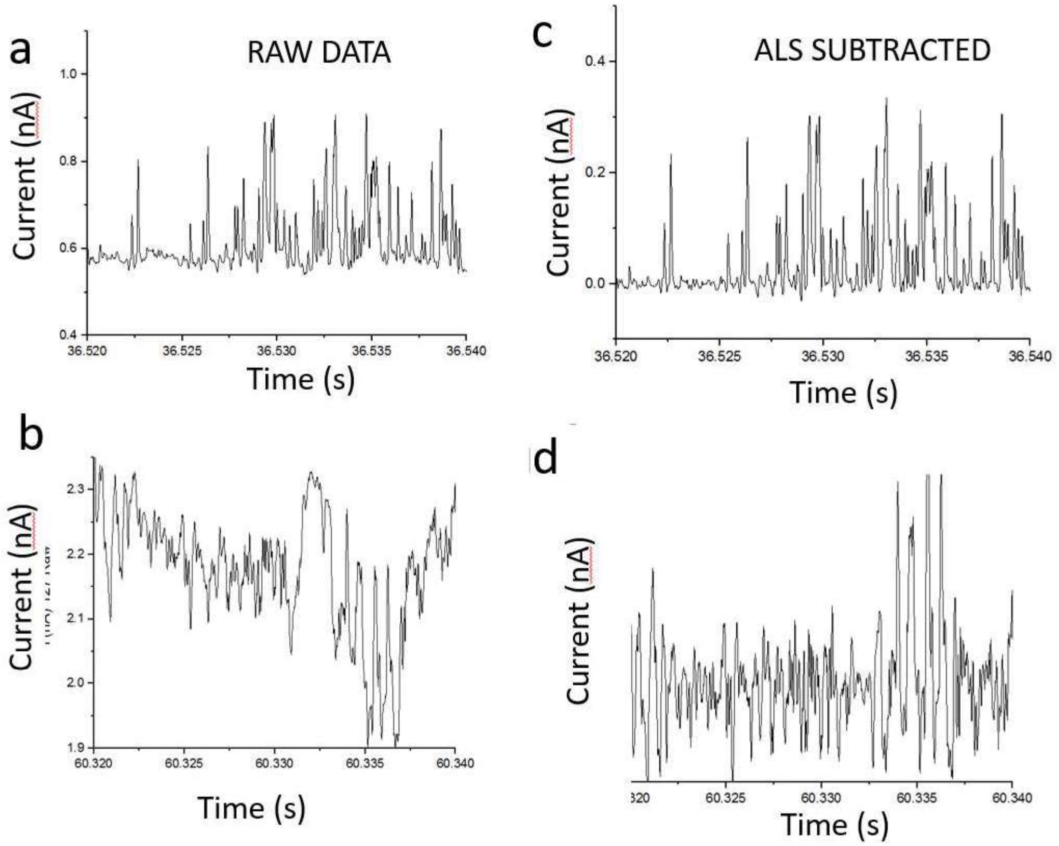


Figure 4.10: Effect of ALS Background Subtraction (Smoothing Factor = 0.1 ms). (a) and (c) are raw and subtracted data for a region of relatively flat baseline. (b) (raw) and (d) (subtracted) illustrate how the noise features are well-preserved even in the presence of large baseline variations.

qualitative observation is complex because the absolute amplitude of the fluctuations depends on the background current, and this current varies during a run. However, the run-to-run variations are generally much larger, so an approximate measure of the relative fluctuation amplitude was obtained as follows: For each molecule measured, ALS fitted baseline currents were binned the as illustrated by the examples in Fig. 4.12 a,c. Many of these distributions could be fitted by a Gaussian (red curve). Many could not; for instance, the background can jump between two or three levels. In these cases, the largest peak that was clear of the background was fitted. The peak of the fitted Gaussian, $I_p$, was then used to characterize the baseline for that run. Examples of the binned noise signals are provided in Fig. 4.12 b,d.

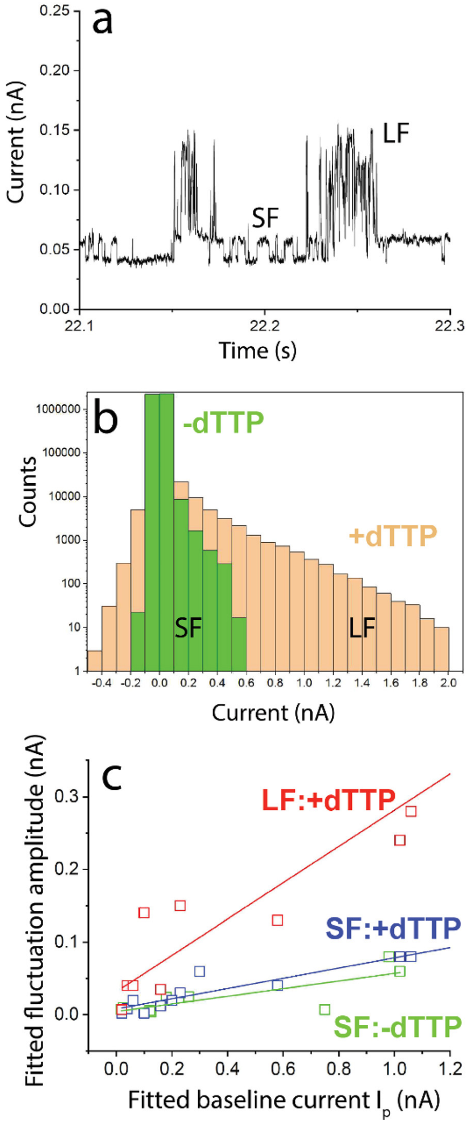


Figure 4.11: Characterizing the Signals Generated When Polymerase Is Activated. (a) Expanded portion of the noise signal measured from an activated polymerase ($dA_{10}$ template) showing two distinct noise components - large fluctuations (LF) and small fluctuations (SF). (b) Distributions of noise signal amplitudes from the active (+dTTP) and inactive (-dTTP) polymerase in runs of about the same background current. (c) Peak noise amplitudes, $I_L$ (LF) and $I_S$ (SF), depend on the baseline current $I_P$. $I_L$ (red) and $I_S$ (blue) are plotted vs the associated value of baseline current ($I_P$). Only SF are observed in inactive (-dTTP) polymerases (green). In the presence of dTTP, 13 samples showed SF (blue points) but only 9 showed LF as well, indicating that four molecules were inactive.

To systematically characterize the binned distributions, a double exponential distribution was employed, where one parameter ($i_S$) represents the amplitude distribution of small fluctuations and another parameter ($i_L$) corresponds to the amplitude distribution of large fluctuations,

$$N(i) = A_1 e^{-\frac{i}{i_S}} + A_2 e^{-\frac{i}{i_L}} \tag{4.1}$$

where $i$ denotes the current in a specific bin of the current distribution.

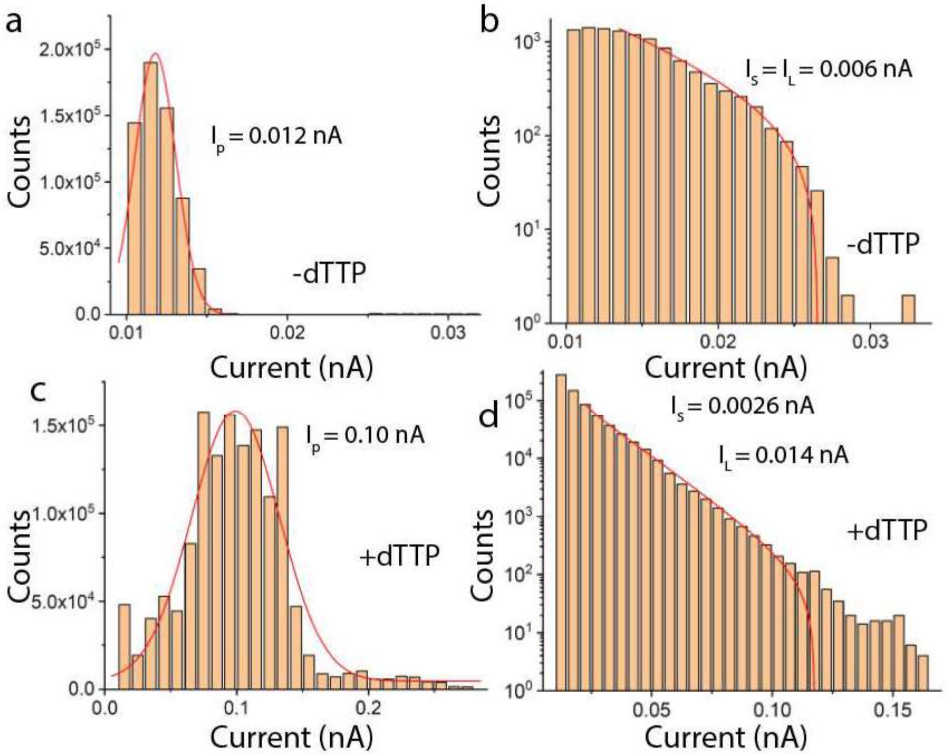


Figure 4.12: Small and Large Fluctuations Defined in Terms of Baseline Current. Larger signals are obtained with the higher - conductance contacts. To quantify this, the baseline levels in a given run are characterized by fitting a Gaussian model to the distribution of measured currents, characterized by the peak value of the Gaussian distribution, $I_P$ (a, c). A two exponential fit ($I_S$, $I_L$) is used to model the distribution of noise signal amplitudes. When the complementary base is absent (b) the fit converges to one value, $I_S = I_L$ in all cases. (d) In the presence of the complementary nucleotide, 9 out of 13 molecules showed a bimodal distribution of spike heights with $I_S << I_L$.

In the experiments without dTTP (e.g., Fig. 4.12 (b)), the fits converged to a single exponential ($i_S = i_L$). For recordings with dTTP, most fits converged on the double-exponential distribution with $i_S < i_L$ (4 out of 13 molecules exhibited small fluctuations only in the presence of dTTP). The results are summarized in Fig. 4.11 c. Activated molecules (+dTTP) primarily exhibited both large (red points) and small (blue points) fluctuations, while the controls (-dTTP) demonstrated only small fluctuations (green points), which were essentially equal to the small fluctuations observed in activated polymerases. A roughly linear relationship between background current and fluctuation amplitude was evidenced by the three linear fits. For the large fluctuations, characteristic of active polymerases, a typical ($1/e$) value of current is $i_L = (0.25 \pm 0.026)i_p$. For the small fluctuations, present in both active and inactive polymerases, a typical value is $i_S = (0.06 \pm 0.01)i_p$. Thus, the active state can be identified by the presence of fluctuations approximately that are about 25% of the baseline current, while fluctuations in the inactive state are about 6% of the baseline current. Not all polymerase molecules contacted were active, as indicated by the absence of large fluctuations in 4 of the 13 molecules studied (these four data points are evident as the extra four blue data points for the SF in the presence of dTTP in Fig. 4.11 c that do not have corresponding red data points for the LF). Conversely, none of the eight -dTTP control runs exhibited large fluctuations (Table 4.2).

This analysis was repeated using data obtained in 38 runs with a d(ATC)5 template (Fig. 4.3 (d)) and 25 runs with a $dC_{10}$ template (Fig. 4.3 (a)). The results are summarized in Fig. 4.13 (a),(b).

The fitted amplitude distributions for the large fluctuations (LF) display considerable variation, but the trends observed for $d(A)_{10}$ (Fig. 4.11 (c)) are well reproduced with $i_L = 0.27(\pm 0.03)I_p$ and $i_S = 0.04(\pm 0.01)I_p$ for $d(ATC)_5$ and $i_L = 0.32(\pm 0.03)I_p$ and $i_S = 0.05(\pm 0.007)I_p$ for $d(C)_{10}$ as shown in Fig. 4.13. In order to confirm the

| Experimental Conditions | Control/Active | Fraction of Large Fluctuations |
|---|---|---|
| $dA_{10}$ + dNTPs + $Mg^{2+}$ | A | 0.69 |
| $dA_{10}$ + dATP + dCTP + dGTP - dTTP + $Mg^{2+}$ | C | 0 |
| $dC_{10}$ + dNTPs + $Mg^{2+}$ | A | 0.52 |
| $(dA_{10})(dC_{10})$ + dNTPs + $Mg^{2+}$ | A | 0.44 |
| $d(ATC)_5$ + dNTPs + $Mg^{2+}$ | A | 0.52 |
| $(dA_{10})(dC_{10})$ - dNTPs + $Mg^{2+}$ | C | 0 |
| $(dA_{10})(dC_{10})$ + dNTPs - $Mg^{2+}$ | C | 0 |
| $(dA_{10})(dC_{10})$ + dTTP - dGTP + $Mg^{2+}$ | C | 0.32 |
| $(dA_{10})(dC_{10})$ + NH-dNTPs + $Mg^{2+}$ | C | 0 |
| no templates + dNTPs + $Mg^{2+}$ | C | 0 |

Table 4.2: Occurrence of Large Fluctuations (Fraction of Measured Molecules) for Various Experimental Conditions. Here, NH = non-hydrolyzable dNTPs. Large fluctuations are identified by a fit converging on a two component distribution of amplitudes.

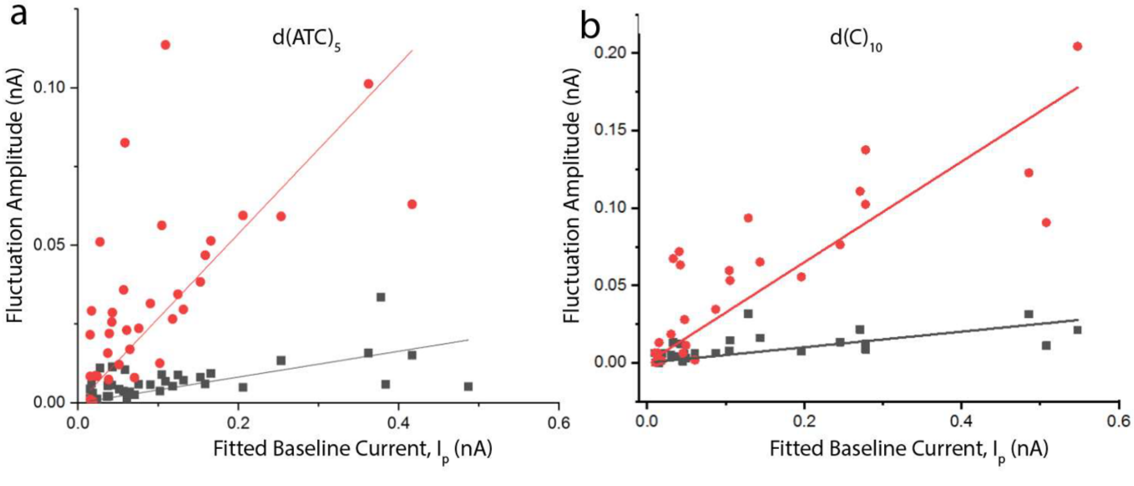


Figure 4.13: Values of $I_L$ (red) and $I_S$ (gray) plotted vs. the associated value of baseline current ($I_P$) for 38 Φ29 molecules actively transcribing the $d(ATC)_5$ template (a) and 25 molecules transcribing the $dC_{10}$ template (b).

association of large fluctuations (amplitudes > 25% of the baseline current) with polymerase activity, experiments were carried out with several DNA templates (Fig. 4.3) under different conditions. Active ("A" in Table 4.2) polymerases were measured in the reaction buffer (1 mM phosphate buffer, pH = 7.4, 4 mM TCEP, 10 mM $MgCl_2$ with 1 mM dNTPs and 1 $\mu$M template). In the control experiments ("C" in Table 4.2), one essential ingredient was withheld. Moreover, measurements employing non-hydrolyzable dNTPs were performed and the fluctuations were analyzed using the

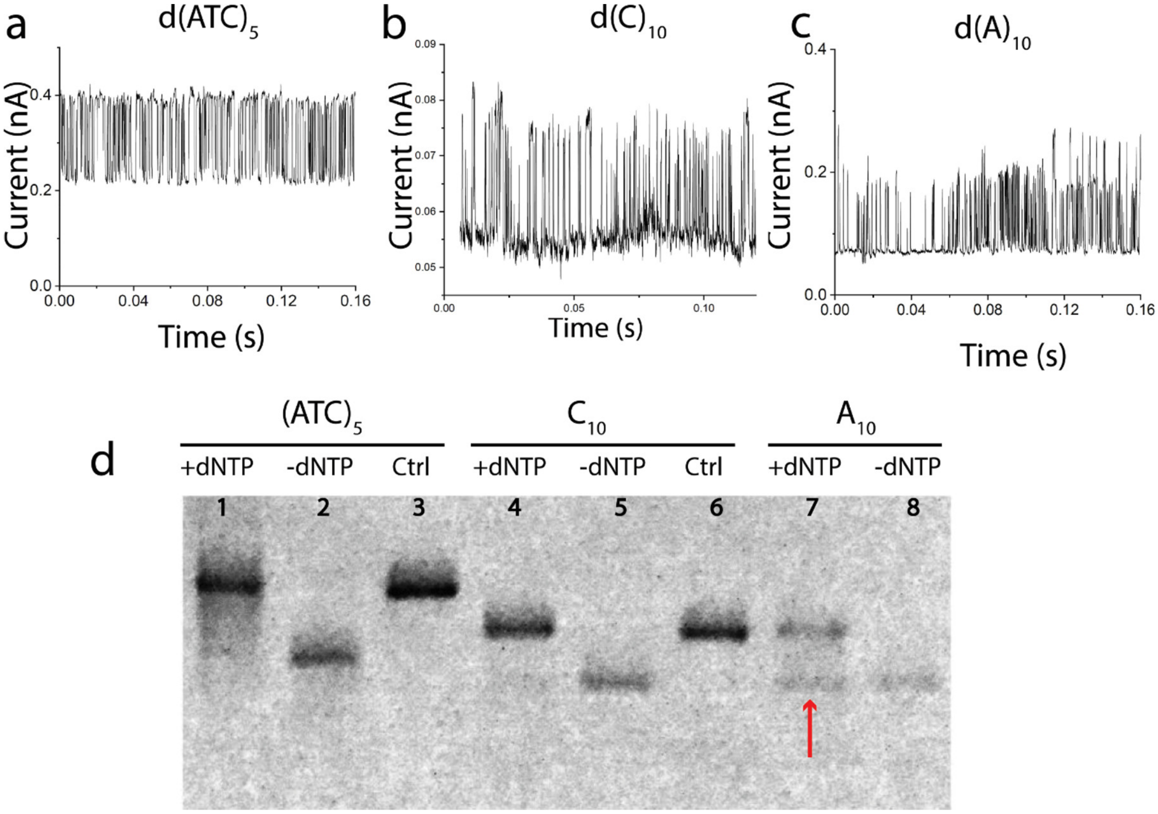
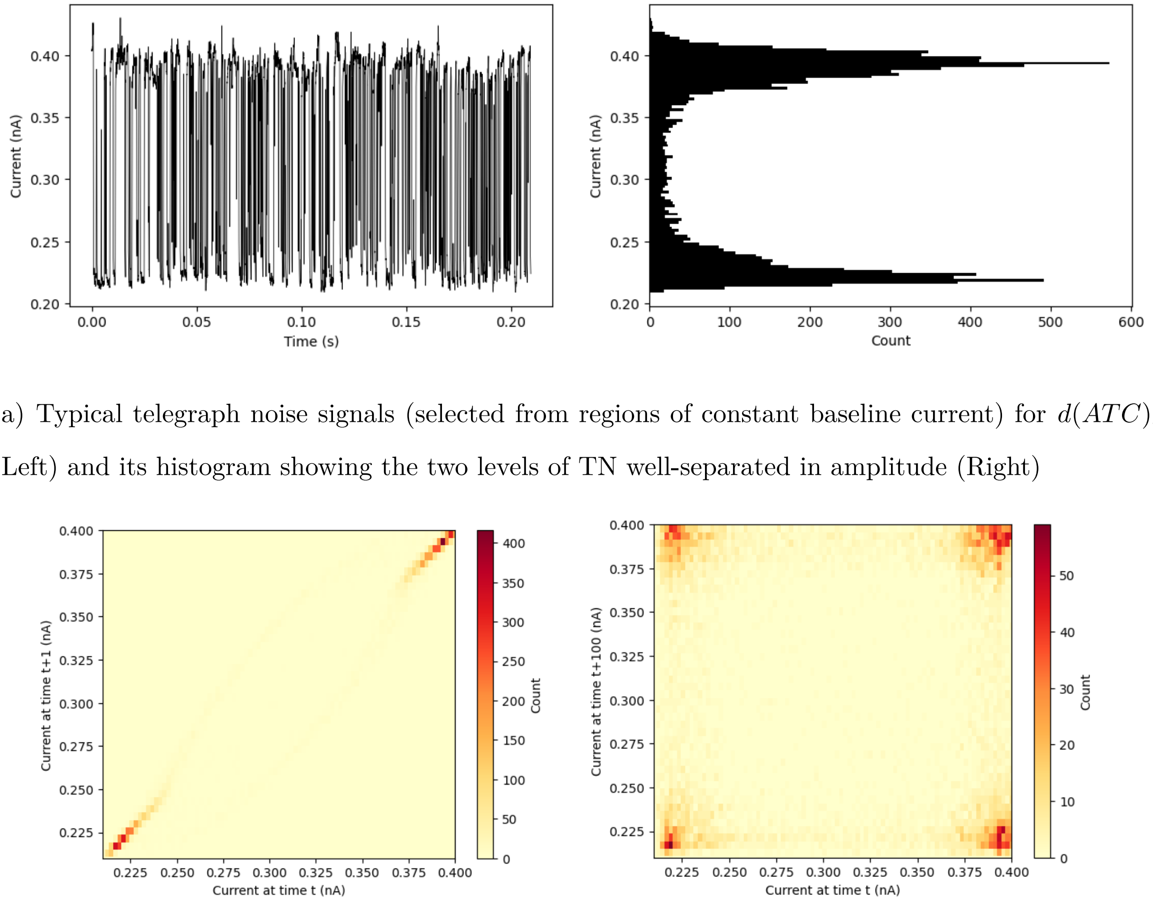


Figure 4.14: Telegraph Noise Signals Reflect the Ease with Which a Template Is Processed. Typical telegraph noise signals (selected from regions of constant baseline current) for $d(ATC)_5$ and $dC_{10}$ are fairly uniform in amplitude and in the period between signal features. Signals for $dA_{10}$ are typically more irregular (c). Denaturing gel (d) shows that the $d(ATC)_5$ and $dC_{10}$ templates are completely converted to full length molecules (the Ctrl lanes are for synthesized fully extended molecules). This is not the case for the $dA_{10}$ (red arrow) for which polymerization is only partially successful ($C_{10}$ control serves as a length control for this molecule also).

previously - mentioned method. It is evident that removing any crucial component for polymerase activity eliminates significant fluctuations. However, an intriguing exception was observed for $(dA_{10})(dC_{10})$ in the presence of dTTP alone. The initial expectation was that the polymerase would proceed to the end of the A tract and subsequently stall due to the absence of the required dGTP nucleotide. Surprisingly, approximately one-third of the molecules seemed to be active during a portion of the recording. It was observed that the telegraph noise acquired from the $(ATC)_5$ and $C_{10}$ templates was typically more regular in both time and amplitude compared to the $A_{10}$ template (Fig. 4.14 (a)–(c)). Furthermore, a denaturing gel displaying the polymerization products (Fig. 4.14 (d)) clearly exhibits incomplete transcription of the $A_{10}$ template (red arrow, lane 7). Consequently, it appears probable that the polymerase dissociates from the A homopolymer tract, enabling another template to bind and accounting for the observed activity in the $A_{10}C_{10}$ template in the absence of dGTP.

Additionally, an estimation of the duration for each conformation of the polymerase can be approximated using a time-lag plot. Time-lag plots, also known as delay or lag plots, are graphical representations of time series data plotted against a version of itself that is lagged by a specific time interval. The $k^{th}$ lag is the time period that happened “k” timestamps after initial timestamp. Fig. 4.15(a) revisits the trace shown in Fig. 4.14 (a) for $d(ATC)_5$ and computes a histogram to clearly show the demarcation of two amplitude levels in the signal. Fig. 4.15(b) illustrates the first-order time lag corresponding to 1 timestamp or 0.00002 s ( the $1^{st}$ lag corresponds to $\frac{1}{50000}$ s, given that the sampling frequency is 50 kHz) for the trace shown in Fig. 4.15(a) for $d(ATC)_5$. Fig. 4.15(c) presents the hundredth-order time lag plot corresponding to 100 timestamps or 2 ms ( the $100^{th}$ lag corresponds to $\frac{100}{50000}$ s, given that the sampling frequency is 50 kHz) for the same trace. It is evident that the two levels of the TN signal are distinctly separated after the $1^{st}$ lag, while the $100^{th}$ introduces two additional clusters, indicating the transition from one level to another. This results in an average dwell time of 2 ms for both levels . Section 4.4.2 provides a more precise estimation of the dwell times of each level in the TN signal.



(a) Typical telegraph noise signals (selected from regions of constant baseline current) for $d(ATC)_5$ (Left) and its histogram showing the two levels of TN well-separated in amplitude (Right)

(b) $1^{st}$-order time-lag plot

(c) $100^{th}$-order time-lag plot

Figure 4.15: Time-lag Plots Estimate the Dwell times of Different Levels in a Typical Telegraph Noise Signal. (a) Typical telegraph noise signals (selected from regions of constant baseline current) for $d(ATC)_5$ (Left) and its histogram showing the two levels of TN well - separated in amplitude (Right). (b) Time-lag plot for lag = 1 timestamp showing the clear separation of two clusters supporting the histogram in (a). (c) Time-lag plot for lag = 100 timestamps showing two additional clusters reflecting the initiation of the transition events between the two levels in the TN signal. This helps in getting an estimate of the average dwell times of the levels in the TN signal.

## 4.4 Current vs. Time (I(t)) Measurements Using a Solid-state Device

This section presents the methodology, experimental results and data interpretation for single molecule conductance measurements carried out using a solid-state device.

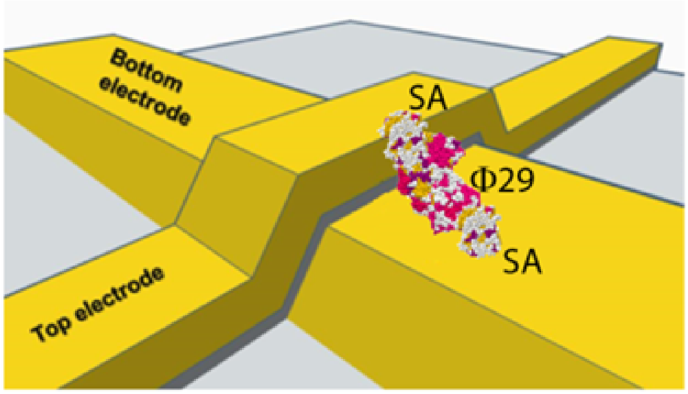


Figure 4.16: Solid-state Device for Stable I(t) Measurements.

### 4.4.1 Experimental Design

Incorporating the system into a solid-state tunnel junction device allows for data collection over long period of time, as well as for the recording of current changes in response to chemical changes. A fixed-gap tunneling device (schematic shown in Fig. 4.16) offers several notable advantages over STM measurements. Firstly, the gap remains constant regardless of electrical operating conditions, allowing the bias to be adjusted without altering the gap. Secondly, the gap can be accurately determined using TEM measurements. Thirdly, the smaller size of the device results in reduced electrode capacitances and an improved frequency response. Finally, when the electrode dimensions are sufficiently small, the simultaneous binding of numerous molecules becomes unlikely, and the stable electrical properties of the junction enable the detection of single-molecule binding events through the production of distinct two-level signals (2LSs). A solid-state device (details of fabrication are described in Section 3.2), was wire-bonded with the chip carrier. The sample was added to a Polydimethylsiloxane (PDMS) cell mounted on the chip. The I(t) data was collected by integrating the chip with the electrical circuitry and the data acquisition system.

### *4.4.2 Dwell Time Measurements*

A solid-state device (Fig. 4.16) was used to monitor rapid polymerase activity over longer periods than possible with the STM by measuring the current as a function of time ($I(t)$) at a bias 50 mV (sampling frequency = 50 kHz) in the presence of template-bound Φ29, $Mg^{2+}$, and dNTPs. This experiment was performed for the $(ATC)_5$ and $A_{10}$ templates. Approximate estimation of the dwell times of the conformations can be done using time-lag plots as shown in Section 4.3.2.4. To accurately estimate the dwell times of the different conformations of the active polymerase, a three-tier algorithm was developed to distinguish between the 'active polymerase' regions and the 'inactive' polymerase regions within the entire signal (Fig. 4.17 shows the different levels of the current amplitude and their significance). This algorithm not only identified the regions of interest (ROIs) from the entire I(t) trace but also facilitated the determination of dwell times for the active polymerase within each Region of Interest (ROI).

The three-step algorithm involved the following procedures (Fig. 4.18):

1. **Application of the cubic splines method:** This mathematical technique, which fits piecewise polynomial functions to a dataset, was used to approximate the lower envelope of the signal, generating a smooth curve that captures the data's essential features. A smoothing window with a length equal to the signal length divided by 1000 was chosen for the lower envelope computation (Fig. 4.18 (a)).

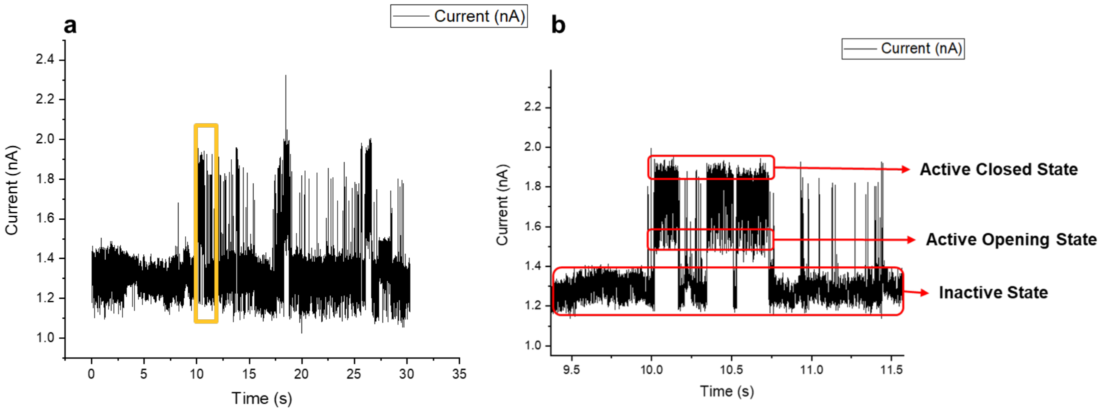


Figure 4.17: Real-time Polymerase Activity: Φ29, $dA_{10}$ Template, dTTP at 50mV Bias. (a) Entire I(t) trace, (b) Zoomed-in section (yellow box in (a)) shows different levels of current amplitude.

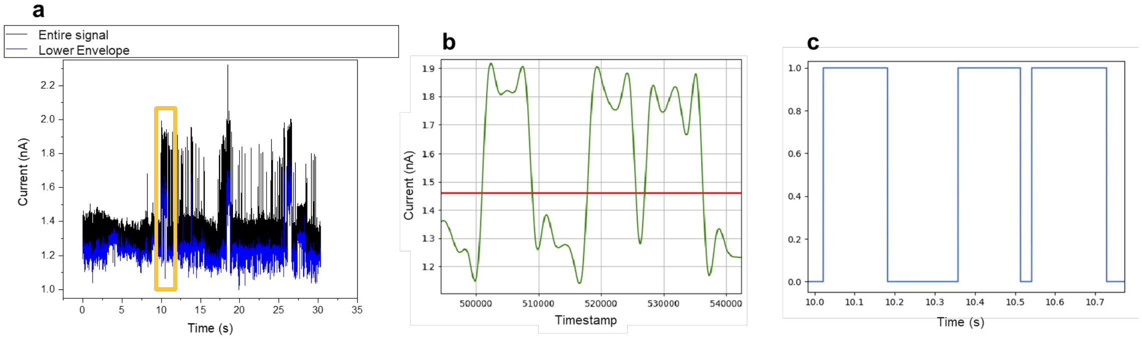


Figure 4.18: A Custom 3-step Algorithm to Separate the Active Polymerase Region from an Entire I(t) Trace. The steps of this Region of Interest (ROI) selector are shown in (a) Application of the cubic splines method to find the lower envelope of the signal (black shows the data, blue shows the lower envelope, yellow box denotes the zoomed-in section shown in the following steps), (b) Thresholding the envelope at 75% of its maximum value, and (c) Converting the envelope to a bi-level square wave to identify the timestamps of the high-current amplitude regions.

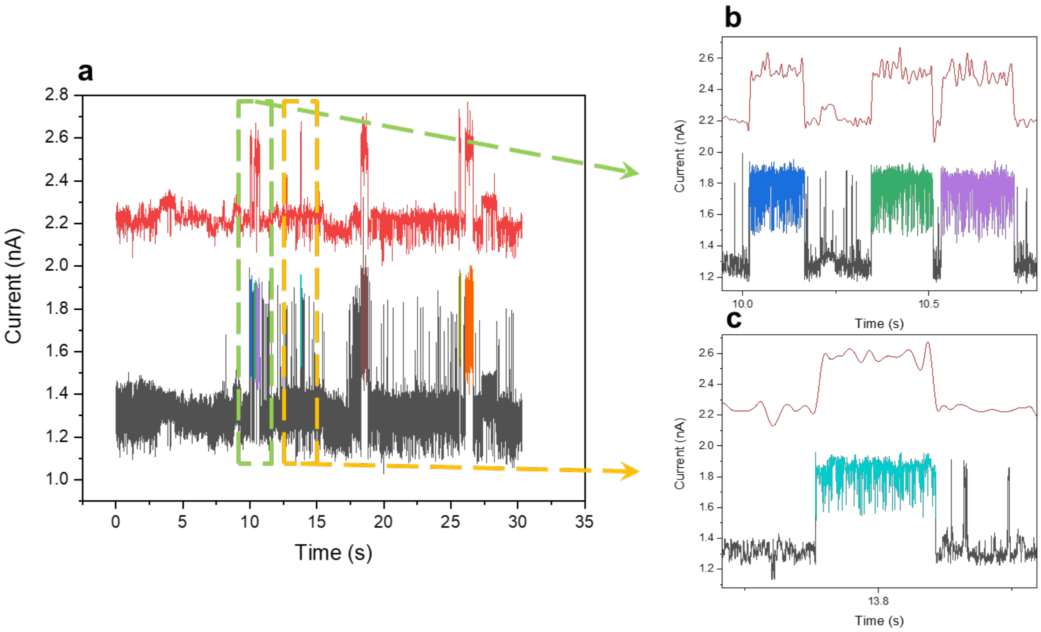


Figure 4.19: Outcome of the Region of Interest (ROI) Selection Algorithm. (a) Entire I(t) Trace (black shows the data, red shows the lower envelope, the green and yellow boxes are explored in (b) and (c)), (b) and (c) highlight the regions selected by the algorithm ((black shows the data, red shows the lower envelope, the blue, green, purple and cyan regions denote the active regions of the polymerase when it continues the DNA synthesis process). This algorithm can be further used on the selected regions to compute dwell times at different current amplitude levels.

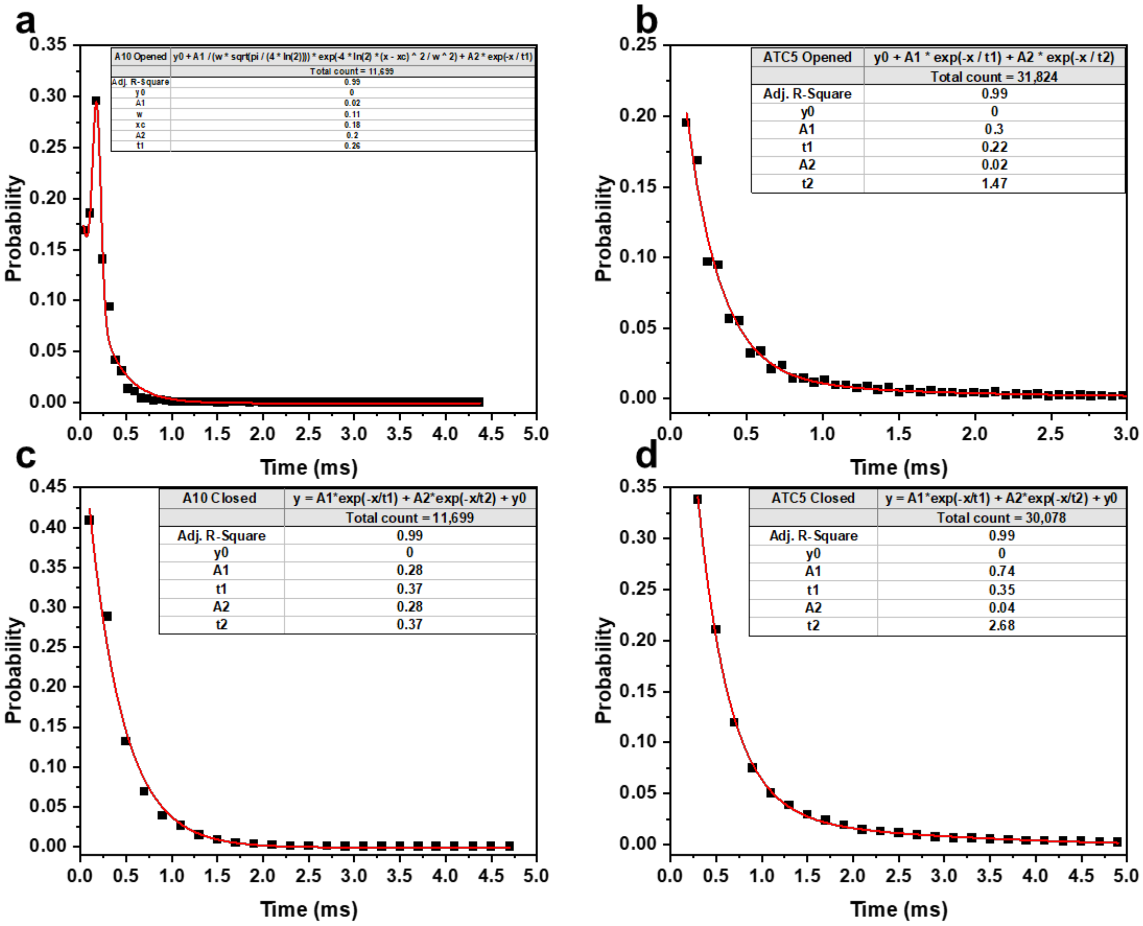
2. **Thresholding the envelope:** The envelope was thresholded at 75% of its maximum value. Data points above the threshold were deemed to represent the '1' level (active polymerase), while points below the threshold were considered as the '0' level (inactive polymerase). This step produced a bi-level square wave comprising 0 and 1 states (Fig. 4.18 (b) and (c)).

3. **Identifying high-current regions:** The timestamps corresponding to the '1' state were used to pinpoint high-current regions, effectively separating the active region of the polymerase from the inactive region in the signal. An ROI was defined if the selected region contained 5000 or more data points, corresponding to a duration of 100 ms (Fig. 4.19).

The ROIs obtained through this algorithm (Fig. 4.19) were preserved for further dwell time analysis, offering valuable insights into the behavior of the activated polymerase. The same ROI-selection algorithm was applied to the ROIs to estimate the 'open' and 'close' dwell times of the activated polymerase (Fig. 4.20). In this context, the '1' state of the bi-level square wave represented the 'close' state of the polymerase, during which the polymerase processes the dNTPs. Conversely, the '0' state corresponded to the 'open' state of the polymerase, which was associated with the capture time of the polymerase. This approach facilitated a comprehensive understanding of the dynamic behavior of the activated polymerase under various conditions.

The analysis of the opening times of the polymerase showed a $\sim$ 7x increase in the speed of dNTP capture in the case of $A_{10}$ (histogram bin centers and Gaussian fit shown in Fig. 4.20 (a)) as opposed to $(ATC)_5$ (histogram bin centers and bi-exponential fit shown in Fig. 4.20 (b)). Fig. 4.20 shows an intrinsic capture time of 0.2 ms for both the templates with an additional higher proofreading time of 1.5 ms in the case of $(ATC)_5$. It was also noticed that 25% of the events in $(ATC)_5$ had the same capture time as that of $A_{10}$.

The analysis of the closed times of the polymerase showed a $\sim$ 8x increase in the speed of dNTP processing in the case of $A_{10}$ (histogram bin centers and bi-exponential fit shown in Fig. 4.20 (c)) as opposed to $(ATC)_5$ (histogram bin centers and bi-exponential fit shown in Fig. 4.20 (d)). Fig. 4.20 shows a processing time of 0.35 ms for both the templates with an additional higher processing time of 2.7 ms in the case of $(ATC)_5$. It was also noticed that 67% of the events in $(ATC)_5$ had the same processing time as that of $A_{10}$. Future studies on opening and closing dwell times can help classify between different templates while understanding the polymerase activity quantitatively.



| A10 Opened | y0 + A1 / (w * sqrt(pi / (4 * ln(2)))) * exp(-4 * ln(2) * (x - xc) ^ 2 / w ^ 2) + A2 * exp(-x / t1) |
|---|---|
| | Total count = 11,699 |
| Adj. R-Square | 0.99 |
| y0 | 0 |
| A1 | 0.02 |
| w | 0.11 |
| xc | 0.18 |
| A2 | 0.2 |
| t1 | 0.26 |

| ATC5 Opened | y0 + A1 * exp(-x / t1) + A2 * exp(-x / t2) |
|---|---|
| | Total count = 31,824 |
| Adj. R-Square | 0.99 |
| y0 | 0 |
| A1 | 0.3 |
| t1 | 0.22 |
| A2 | 0.02 |
| t2 | 1.47 |

| A10 Closed | y = A1*exp(-x/t1) + A2*exp(-x/t2) + y0 |
|---|---|
| | Total count = 11,699 |
| Adj. R-Square | 0.99 |
| y0 | 0 |
| A1 | 0.28 |
| t1 | 0.37 |
| A2 | 0.28 |
| t2 | 0.37 |

| ATC5 Closed | y = A1*exp(-x/t1) + A2*exp(-x/t2) + y0 |
|---|---|
| | Total count = 30,078 |
| Adj. R-Square | 0.99 |
| y0 | 0 |
| A1 | 0.74 |
| t1 | 0.35 |
| A2 | 0.04 |
| t2 | 2.68 |

Figure 4.20: Opening and Closing Times of Chip-mounted Polymerase Estimated Using a 3-step Algorithm. (a) and (b) show the Gaussian and Bi-exponential fits for 'Open' times of the polymerase in the presence of $dA_{10}$ and $d(ATC)_5$ templates respectively. 'N' denotes the total number of samples and the scatter points are the histogram bin centers. (c) and (d) show the Bi-exponential fits for 'Closed' times of the polymerase in the presence of $dA_{10}$ and $d(ATC)_5$ templates respectively.

### *4.4.3 Repeating Motifs in the Current Signature*

#### *4.4.3.1 Circular DNA Template*

For this experiment, a linear single-stranded oligonucleotide RCR ($5'$- p - CATCTACTACGCTTAGCTTGCTATCATCTATGCTTAGCATGC - $3'$) was employed to generate a circular RCR template through enzymatic self-ligation using Circligase (Epicentre). In a 1X reaction buffer containing 50 $\mu$M ATP and 2.5 mM $MnCl_2$, 0.1 nmol of linear single-stranded RCR DNA was combined with 100 Units of Circligase. Following incubation at 60°C for 2 hours, the product was heated to 80°C for 10 minutes in order to inactivate the Circligase. Linear ssDNA remaining in the solution was digested by Exo I (NEB). Quality control of the RCR template was performed through electrophoresis on a denaturing gel containing 8 M urea and 20% polyacrylamide. For later use, 2.5 pmols of RCR template was annealed with 50 pmols of a 21-mer RCR primer (5'-GGCATGCTAAGCATAGATGAT-3') by heating to 95°C for 5 minutes and gradually cooling down to room temperature (decreasing 0.1°C/s) and stored at -20°C.

#### *4.4.3.2 Results and Discussion*

As described in Section 1.3.2, matrix profile helps in motif discovery or finding repeating patterns in the I(t) trace. A time series motif is a group of well-conserved subsequences (patterns) in a time series [21]. A matrix profile was computed for a signal ROI (Fig. 4.21 (c)) from the solid-state device-aided I(t) measurements for the 42-nucleotide circular DNA template (sequence described in Section 4.3.1.3). Fig. 4.21 (d) presents a matrix profile, where relatively low values indicate that the subsequence in the original time series must have (at least one) relatively similar subsequence(pattern) elsewhere in the data (such regions are “motifs” or recurring patterns). Different window sizes were employed to find 4 different motifs (assuming that the 4 motifs corresponded to 4 different nucleotides, namely ‘A’, ‘T’, ‘G’ and ‘C’) with no overlap (Fig. 4.22). The window was optimized to be 6144 data points. The length of the ROI was 130229 data points. This essentially meant that the ROI was segregated into  21 segments ($= \frac{130229}{6144}$), which was exactly half of the DNA template length.

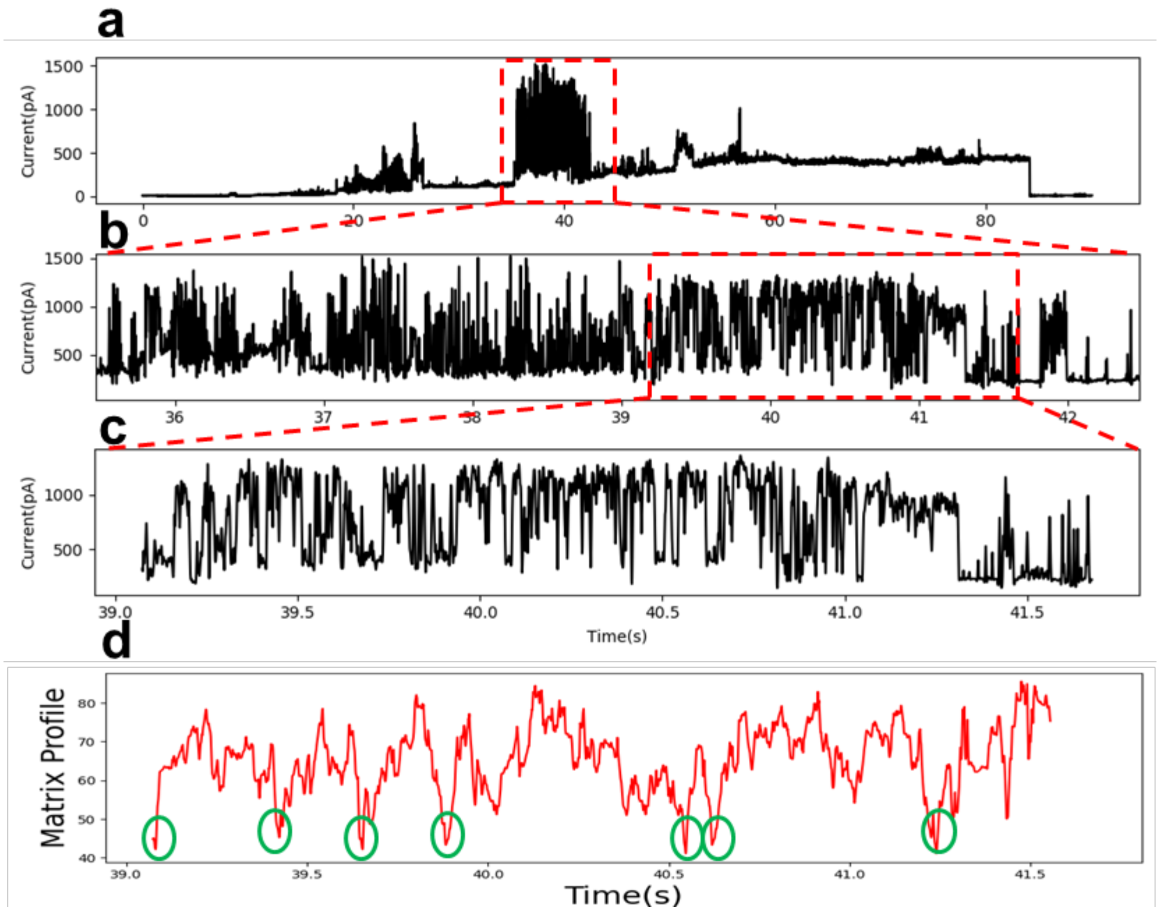


Figure 4.21: A Matrix Profile Was Computed for the I(t) Trace Collected Using a Solid-state Device in the Presence of Φ29, a 42-Nucleotide Circular DNA Template and dNTPs. (a) Entire I(t) Trace, (b) ∼8s Stable (Drift-Free) Amplitude Region Selected from ALS Corrected Signal, (c) The TN dominating region of the ROI to be used for motif discovery. (d) The matrix profile computed for the signal in (c). In a matrix profile, relatively low values indicate the beginning of a “motif” or a recurring pattern (marked in green circles)

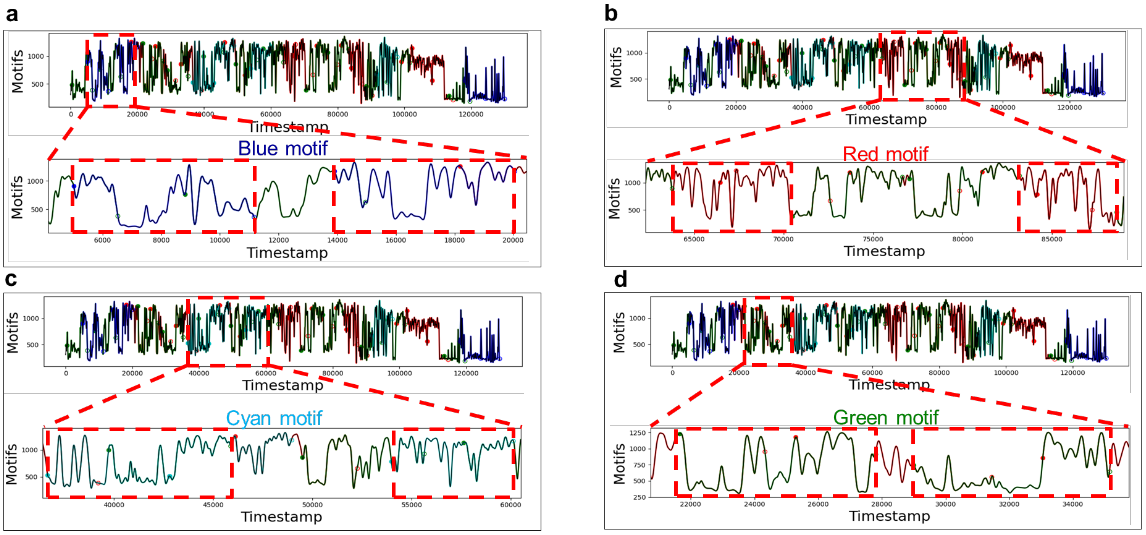


Figure 4.22: The Motif (Subsequence) Structures Derived from the Matrix Profile Values Are Significantly Different. Panels (a),(b),(c) and (d) show the blue , red, cyan and green motifs respectively (assuming that the 4 motifs corresponded to 4 different nucleotides, namely 'A', 'T', 'G' and 'C').

Further experiments are needed to support the findings of matrix profiling. This study provided a way to isolate specific motifs for each nucleotide. As seen in the Fig. 4.22, the motif structures are significantly different. The motif statistics (for example, the number of peaks and the width of peaks) can be studied and further employed in a machine learning algorithm that can recognize the same motifs from the current/noise signature of an unknown template.

## 4.5 Conclusion

This chapter demonstrates that enzyme activity can be monitored through direct electrical measurements, paving the way for incorporating the analytical power of enzymes into integrated circuits, provided further studies on integrating the molecules into solid-state gap devices are performed [19]. Engineering two contact points into a polymerase results in conductance distribution features that are approximately 3–10 times larger than those observed with only one engineered contact and a second, nonspecific contact. The conductance of the complex formed by streptavidin and doubly biotinylated Φ29 is further increased if biotin is used to anchor the streptavidin to the electrodes rather than thiolated surface lysines. Significant changes in the conductivity distribution occur as the polymerase undergoes the open-to-closed transition. Moreover, polymerase activity is characterized by rapid noise spikes with an amplitude of approximately 25% (or more) of the background current, which is distinct from the smaller (6% of background) signals present in both active and inactive polymerase. A deeper understanding of these signals necessitates investigating molecules wired into solid-state gap devices, for which much longer and less variable data runs can be obtained. Fixed junction devices have been created by drilling an orifice through a stack of metal-dielectric-metal layers [22], and the adaptation of these devices as fixed contacts for enzyme measurements is being explored. Such devices yield data free from interruptions due to contact drift, enabling the determination of whether the pauses observed in the signals (Fig. 4.8 (b)) are intrinsic to the polymerase or not. Preliminary results from such devices show a stable contact allowed the collection of much longer data trains so that pauses induced by slower binding of dNTPs and template could be investigated by analyzing the dwell times of the different current amplitude levels.

# REFERENCES

[1] Bintian Zhang, Hanqing Deng, Sohini Mukherjee, Weisi Song, Xu Wang, and Stuart Lindsay. Engineering an Enzyme for Direct Electrical Monitoring of Activity. *ACS Nano*, 14(2):1360–1368, 2020.

[2] B. Zhang and S. Lindsay. Electronic Decay Length in a Protein Molecule. *Nano Lett.*, 19:4017, 2019.

[3] Bintian Zhang, Weisi Song, Pei Pang, Huafang Lai, Qiang Chen, Peiming Zhang, and Stuart Lindsay. Role of contacts in long-range protein conductance. *Proceedings of the National Academy of Sciences*, 116(13):5886–5891, 2019.

[4] G. Vattay, D. Salahub, I. a. Csabai, A. Nassimi, and S. A. Kaufmann. Quantum Criticality at the Origin of Life. *J. Phys.: Conf. Ser.*, 626:012023, 2015.

[5] L. Blanco, A. Bernad, J. M. Lazaro, G. Martin, C. Garmendia, and M. Salas. Highly Efficient DNA Synthesis by the Phage Phi 29 DNA Polymerase: Symmetrical Mode of DNA Replication. *J. Biol. Chem.*, 264:8935, 1989.

[6] L. Blanco and M. Salas. Relating Structure to Function in Phi29 DNA Polymerase. *J. Biol. Chem.*, 271:8509, 1996.

[7] J. A. Esteban, M. Salas, and L. Blanco. Fidelity of Phi 29 DNA Polymerase. Comparison Between Protein-Primed Initiation and DNA Polymerization. *J. Biol. Chem.*, 268:2719, 1993.

[8] J. Korlach, A. Bibillo, J. Wegener, P. Peluso, T. T. Pham, I. Park, S. Clark, G. A. Otto, and S. W. Turner. Long, Processive Enzymatic DNA Synthesis Using 100% Dye-Labeled Terminal Phosphate-Linked Nucleotides. *Nucleosides, Nucleotides Nucleic Acids*, 27:1072, 2008.

[9] T. A. Steitz. A Mechanism for All Polymerases. *Nature*, 391:231, 1998.

[10] M. Fairhead and M. Howarth. Site-Specific Biotinylation of Purified Proteins Using BirA. *Methods Mol. Biol.*, 1266:171, 2015.

[11] A. J. Berman, S. Kamtekar, J. L. Goodman, J. M. Lazaro, M. de Vega, L. Blanco, M. Salas, and T. A. Steitz. Structures of Phi29 DNA Polymerase Complexed with Substrate: the Mechanism of Translocation in B-family Polymerases. *EMBO J.*, 26:3494, 2007.

[12] S. Kamtekar, A. J. Berman, J. Wang, J. M. Lazaro, M. de Vega, L. Blanco, M. Salas, and T. A. Steitz. Insights into Strand Displacement and Processivity from the Crystal Structure of the Protein-Primed DNA Polymerase of Bacteriophage Phi29. *Mol. Cell*, 16:609, 2004.

[13] Michael Tuchband, Jin He, Shuo Huang, and Stuart Lindsay. Insulated gold scanning tunneling microscopy probes for recognition tunneling in an aqueous environment. *Review of Scientific Instruments*, 83(1):015102, 2012.

[14] P. Ghisellini, M. Caiazzo, A. Alessandrini, R. Eggenhoffner, M. Vassalli, and P. Facci. Direct Electrical Control of IgG Conformation and Functional Activity at Surfaces. *Sci. Rep.*, 6:37779, 2016.

[15] D. Beckett, E. Kovaleva, and P. J. Schatz. A Minimal Peptide Substrate in Biotin Holoenzyme Synthetase-Catalyzed Biotinylation. *Protein Sci.*, 8:921, 1999.

[16] W. A. Beard, D. D. Shock, V. K. Batra, L. C. Pedersen, and S. H. Wilson. DNA Polymerase Beta Substrate Specificity: Side Chain Modulation of the "A - Rule". *J. Biol. Chem.*, 284:31680, 2009.

[17] J. A. Morin, F. J. Cao, J. M. Lazaro, J. R. Arias-Gonzalez, J. M. Valpuesta, J. L. Carrascosa, M. Salas, and B. Ibarra. Mechano-Chemical Kinetics of DNA Replication: Identification of the Translocation Step of a Replicative DNA Polymerase. *Nucleic Acids Res.*, 43:3643, 2015.

[18] J. M. Lazaro, L. Blanco, and M. Salas. Purification of Bacteriophage Phi 29 DNA Polymerase. *Methods Enzymol.*, 262:42, 1995.

[19] B. Zhang, W. Song, P. Pang, Y. Zhao, P. Zhang, I. Csabai, G. Vattay, and S. Lindsay. Observation of Giant Conductance Fluctuations in a Protein. *Nano Futures*, 1:035002, 2017.

[20] J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, and J. Tan. Asymmetric Least Squares for Multiple Spectra Baseline Correction. *Anal. Chim. Acta*, 683:63, 2010.

[21] Andrew Van Benschoten, Austin Ouyang, Francisco Bischoff, and Tyler Marrs. MPA: a novel cross-language API for time series analysis. *Journal of Open Source Software*, 5(49):2179, 2020.

[22] P. Pang, B. A. Ashcroft, W. Song, P. Zhang, S. Biswas, Q. Qing, J. Yang, R. J. Nemanich, J. Bai, J. T. Smith, K. Reuter, V. S. K. Balagurusamy, Y. Astier, G. Stolovitzky, and S. Lindsay. Fixed-Gap Tunnel Junction for Reading DNA Nucleotides. *ACS Nano*, 8:11994, 2014.

# Chapter 5

# FUTURE DIRECTIONS

Proteins are incredibly versatile and serve as excellent candidates for integration into bioelectronic devices due to their performance in signal transduction, molecular recognition, and selective catalysis. However, they are often considered insulating. The previous chapters presented the research on charge transport in non-redox-active proteins using Scanning tunneling microscopy (STM). STM is a powerful technique that has revolutionized our understanding of molecular systems at the atomic scale. The examples discussed in the thesis serve as proof of concept in laboratory research, giving a chance to realize practical single-molecule sensing devices by overcoming the challenges of the complexity of the two-terminal architecture in STM and constructing robust, reproducible junctions for market use. This chapter will highlight the future possibilities of using single-molecule non-redox protein conductance measurements with STM in applications such as DNA sequencing, single-molecule protein sequencing, antibody-based sensors, and more. Traditional DNA sequencing methods often require amplification and labeling steps, which can introduce errors and biases. STM techniques can overcome these limitations by directly detecting individual DNA nucleotides as they are polymerized by a DNA polymerase, thereby altering patterns in the resulting electrical signature. This approach holds promise for faster, more accurate DNA sequencing with real-time analysis capabilities. Another area of interest is the development of single-molecule protein sequencing techniques that combine STM with protease enzymes. In this context, a protease enzyme cleaves a protein into smaller peptides, which are then identified by their conductance properties. This method could offer an alternative to mass spectrometry for protein identification and quantification. Furthermore, it could enable the study of post-translational modifications and protein-protein interactions in greater detail and facilitate the discovery of novel biomarkers for disease diagnosis and monitoring, as well as the identification of potential drug targets. Single-molecule conductance measurements with STM can also be utilized to develop highly sensitive and specific immunosensors. An immunosensor is a type of biosensor in which a specific target analyte, antigen (Ag), is detected by the formation of a stable immunocomplex between antigen and antibody as a capture agent (Ab). By forming specific contacts to an antibody through a ligand tethered to the STM probe and substrate, researchers can create a selective sensor that responds to a particular target antigen molecule. As the target molecule binds to the antibody, changes in conductance can be detected, allowing for real-time monitoring and quantification of the target. Such sensors hold great potential for the development of point-of-care diagnostics, environmental monitoring, and drug discovery. They could enable the rapid detection of pathogens, toxins, or biomarkers in various samples, including blood, saliva, and water, without the need for complex and time-consuming laboratory procedures. Beyond the applications discussed above, STM-based single-molecule conductance measurements have the potential to revolutionize other areas of molecular biology. For instance, they could be employed to study the folding and dynamics of individual biomolecules or to investigate the temperature or pH-based interactions between molecules in complex cellular environments. Moreover, the integration of this technology with other analytical techniques, such as imaging or spectroscopy, could provide a more comprehensive understanding of biomolecular systems. Advanced nanofabrication techniques could be used to construct more stable molecular devices with high efficiency and scalability, as presented in Chapters 3 and 4. However, such techniques involve challenges such as noisy signals, complex data, and the need for real-time analysis call for developing and applying advanced signal processing techniques and deep learning algorithms. Noise reduction, feature extraction, and data compression are essential for enhancing the signal-to-noise ratio and extracting meaningful information from the raw data. Methods such as wavelet transforms, filtering, and principal component analysis can help preprocess and analyze the complex signals generated during single-molecule DNA and protein sequencing. These techniques can enable the identification of unique conductance signatures corresponding to individual nucleotides or amino acids, facilitating more accurate and efficient sequencing. Deep learning algorithms, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can be employed to recognize complex patterns in the data. CNNs are especially suitable for detecting spatial patterns in STM images, while RNNs can capture temporal dependencies in conductance measurements. Furthermore, the development of specialized deep learning architectures and training strategies tailored to STM data can help optimize performance and reduce computational costs. This will enable researchers to effectively analyze large datasets and make rapid, accurate predictions for DNA and protein sequences. As technology and algorithms advance, we can expect significant breakthroughs in the fields of DNA and protein sequencing, leading to a deeper understanding of biological systems and improved healthcare outcomes.

# REFERENCES

[1] Leandro C Tabares, Ankur Gupta, Thijs J Aartsma, and Gerard W Canters. Tracking electrons in biological macromolecules: From ensemble to single molecule. *Molecules*, 19(8):11660–11678, 2014.

[2] Boris Rotman. Measurement of activity of single molecules of $\beta$-D-galactosidase. *Proceedings of the National Academy of Sciences*, 47(12):1981–1991, 1961.

[3] Ashok A Deniz, Samrat Mukhopadhyay, and Edward A Lemke. Single-molecule biophysics: at the interface of biology, physics and chemistry. *Journal of the Royal Society Interface*, 5(18):15–45, 2008.

[4] Shelley A Claridge, Jeffrey J Schwartz, and Paul S Weiss. Electrons, photons, and force: quantitative single-molecule measurements from physics to biology. *ACS Nano*, 5(2):693–729, 2011.

[5] Massimiliano Di Ventra and Masateru Taniguchi. Decoding DNA, RNA and peptides with quantum tunnelling. *Nature Nanotechnology*, 11(2):117–126, 2016.

[6] William E Moerner and Michel Orrit. Illuminating single molecules in condensed matter. *Science*, 283(5408):1670–1676, 1999.

[7] Ben NG Giepmans, Stephen R Adams, Mark H Ellisman, and Roger Y Tsien. The fluorescent toolbox for assessing protein location and function. *Science*, 312(5771):217–224, 2006.

[8] Joshua Hihath, Bingqian Xu, Peiming Zhang, and Nongjian Tao. Study of single-nucleotide polymorphisms by means of electrical conductance measurements. *Proceedings of the National Academy of Sciences*, 102(47):16979–16983, 2005.

[9] Xiaoyin Xiao, Bingqian Xu, and Nongjian Tao. Changes in the conductance of single peptide molecules upon metal-ion binding. *Angewandte Chemie International Edition*, 43(45):6148–6152, 2004.

[10] Fang Chen, Joshua Hihath, Zhifeng Huang, Xiulan Li, and NJ Tao. Measurement of single-molecule conductance. *Annu. Rev. Phys. Chem.*, 58:535–564, 2007.

[11] Amanda M Moore and Paul S Weiss. Functional and spectroscopic measurements with scanning tunneling microscopy. *Annu. Rev. Anal. Chem.*, 1:857–882, 2008.

[12] Gavin David Scott and Douglas Natelson. Kondo resonances in molecular devices. *ACS Nano*, 4(7):3560–3579, 2010.

[13] T Albrecht. Electrochemical tunnelling sensors and their potential applications. *Nature Communications*, 3(1):829, 2012.

[14] Wikimedia Commons. File:rastertunnelmikroskop-schema.svg — wikimedia commons, the free media repository, 2020. URL `https://commons.wikimedia.org/w/index.php?title=File:Rastertunnelmikroskop-schema.svg&oldid=456882138`.

[15] Francesca Moresco. Manipulation of large molecules by low-temperature STM: model systems for molecular electronics. *Physics Reports*, 399(4):175–225, 2004.

[16] Supriyo Datta, Weidong Tian, Seunghun Hong, R Reifenberger, Jason I Henderson, and Clifford P Kubiak. Current-voltage characteristics of self-assembled monolayers by scanning tunneling microscopy. *Physical Review Letters*, 79(13): 2530, 1997.

[17] Shuai Chang, Jin He, Ashley Kibel, Myeong Lee, Otto Sankey, Peiming Zhang, and Stuart Lindsay. Tunnelling readout of hydrogen-bonding-based recognition. *Nature Nanotechnology*, 4(5):297–301, 2009.

[18] Yanan Zhao, Brian Ashcroft, Peiming Zhang, Hao Liu, Suman Sen, Weisi Song, JongOne Im, Brett Gyarfas, Saikat Manna, Sovan Biswas, et al. Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nature Nanotechnology*, 9(6):466–473, 2014.

[19] Abraham Nitzan. *Chemical dynamics in condensed phases: relaxation, transfer and reactions in condensed molecular systems.* Oxford university press, 2006.

[20] Christopher D Bostick, Sabyasachi Mukhopadhyay, Israel Pecht, Mordechai Sheves, David Cahen, and David Lederman. Protein bioelectronics: a review of what we do and do not know. *Reports on Progress in Physics*, 81(2):026601, 2018.

[21] Ramesh Y Adhikari, Nikhil S Malvankar, Mark T Tuominen, and Derek R Lovley. Conductivity of individual Geobacter pili. *RSC Advances*, 6(10):8354–8357, 2016.

[22] F. J. R. Meysman, R. Cornelissen, S. Trashin, R. Bonne, S. H. Martinez, J. van der Veen, C. J. Blom, C. Karman, J. L. Hou, R. T. Eachambadi, J. S. Geelhoed, K. Wael, H. J. E. Beaumont, B. Cleuren, R. Valcke, H. S. J. van der Zant, H. T. S. Boschker, and J. V. Manca. A Highly Conductive Fibre Network Enables Centimetre-Scale Electron Transport in Multicellular Cable Bacteria. *Nat. Commun.*, 10:4120, 2019.

[23] N. Amdursky, D. Marchak, L. Sepunaru, I. Pecht, M. Sheves, and D. Cahen. Electronic Transport via Proteins. *Adv. Mater.*, 26:7142, 2014.

[24] G. Vattay, D. Salahub, I. a. Csabai, A. Nassimi, and S. A. Kaufmann. Quantum Criticality at the Origin of Life. *J. Phys.: Conf. Ser.*, 626:012023, 2015.

[25] Stuart Lindsay. Ubiquitous Electron Transport in Non-Electron Transfer Proteins. *Life*, 10(5), 2020.

[26] Jesús E Contreras-Naranjo and Oscar Aguilar. Suppressing non-specific binding of proteins onto electrode surfaces in the development of electrochemical immunosensors. *Biosensors*, 9(1):15, 2019.

[27] B. Zhang, W. Song, P. Pang, Y. Zhao, P. Zhang, I. Csabai, G. Vattay, and S. Lindsay. Observation of Giant Conductance Fluctuations in a Protein. *Nano Futures*, 1:035002, 2017.

[28] Bintian Zhang, Weisi Song, Pei Pang, Huafang Lai, Qiang Chen, Peiming Zhang, and Stuart Lindsay. Role of contacts in long-range protein conductance. *Proceedings of the National Academy of Sciences*, 116(13):5886–5891, 2019.

[29] B. Zhang and S. Lindsay. Electronic Decay Length in a Protein Molecule. *Nano Lett.*, 19:4017, 2019.

[30] Bintian Zhang, Weisi Song, Jesse Brown, Robert Nemanich, and Stuart Lindsay. Electronic Conductance Resonance in Non-Redox-Active Proteins. *Journal of the American Chemical Society*, 142(13):6432–6438, 2020.

[31] Adi Salomon, David Cahen, Stuart Lindsay, John Tomfohr, Vincent B Engelkes, and C Daniel Frisbie. Comparison of electronic transport measurements on organic molecules. *Advanced Materials*, 15(22):1881–1890, 2003.

[32] Q. Lu, K. Liu, H. Zhang, Z. Du, X. Wang, and F. Wang. From Tunneling to Hopping: A Comprehensive Investigation of Charge Transport Mechanism in Molecular Junctions Based on Oligo(p-phenylene ethynylene)s. *ACS Nano*, 3: 3861, 2009.

[33] Bintian Zhang, Eathen Ryan, Xu Wang, Weisi Song, and Stuart Lindsay. Electronic Transport in Molecular Wires of Precisely Controlled Length Built from Modular Proteins. *ACS Nano*, 16(1):1671–1680, 2022.

[34] Catharine Shipps, H. Ray Kelly, Peter J. Dahl, Sophia M. Yi, Dennis Vu, David Boyer, Calina Glynn, Michael R. Sawaya, David Eisenberg, Victor S. Batista, and Nikhil S. Malvankar. Intrinsic electronic conductivity of individual atomically resolved amyloid crystals reveals micrometer-long hole hopping via tyrosines. *Proceedings of the National Academy of Sciences*, 118(2):e2014139118, 2021.

[35] L. Sepunaru, I. Pecht, M. Sheves, and D. Cahen. Solid-State Electron Transport across Azurin: From a Temperature-Independent to a Temperature-Activated Mechanism. *J. Am. Chem. Soc.*, 133:2421, 2011.

[36] K. S. Kumar, R. R. Pasula, S. Lim, and C. A. Nijhuis. Long-Range Tunneling Processes across Ferritin-Based Junctions. *Adv. Mater.*, 28:1824, 2016.

[37] B. Kayser, J. A. Fereiro, R. Bhattacharyya, S. R. Cohen, A. Vilan, I. Pecht, M. Sheves, and D. Cahen. Solid-State Electron Transport via the Protein Azurin is Temperature-Independent Down to 4 K. *journal of Physical Chemistry Letters*, 11:144, 2020.

[38] S. Mukhopadhyay, S. Dutta, I. Pecht, M. Sheves, and D. Cahen. Conjugated Cofactor Enables Efficient Temperature-Independent Electronic Transport Across approximately 6 nm Long Halorhodopsin. *J. Am. Chem. Soc.*, 137:11226, 2015.

[39] K. Garg, M. Ghosh, T. Eliash, J. H. van Wonderen, J. N. Butt, L. Shi, X. Jiang, F. Zdenek, J. Blumberger, I. Pecht, M. Sheves, and D. Cahen. Direct evidence for heme-assisted solid-state electronic conduction in multi-heme c-type cytochromes. *Chem. Sci.*, 9:7304, 2018.

[40] K. Garg, S. Raichlin, T. Bendikov, I. Pecht, M. Sheves, and D. Cahen. Interface Electrostatics Dictates the Electron Transport via Bioelectronic Junctions. *ACS Appl. Mater. Interfaces*, 10:41599, 2018.

[41] J. A. Fereiro, X. Yu, I. Pecht, M. Sheves, J. C. Cuevas, and D. Cahen. Tunneling explains efficient electron transport via protein junctions. *Proc. Natl. Acad. Sci. U. S. A.*, 115:E4577, 2018.

[42] O. E. Castañeda Ocampo, P. Gordiichuk, S. Catarci, D. A. Gautier, A. Herrmann, and R. C. Chiechi. Mechanism of Orientation-Dependent Asymmetric Charge Transport in Tunneling Junctions Comprising Photosystem I. *J. Am. Chem. Soc.*, 137:8419, 2015.

[43] Q. Chi, O. Farver, and J. Ulstrup. Long-range protein electron transfer observed at the single-molecule level: In situ mapping of redox-gated tunneling resonance. *Proc. Natl. Acad. Sci. U. S. A.*, 102:16203, 2005.

[44] A. Alessandrini, S. Corni, and P. Facci. Unravelling single metalloprotein electron transfer by scanning probe techniques. *Phys. Chem. Chem. Phys.*, 8:4383, 2006.

[45] E. A. Pia, Q. Chi, D. D. Jones, J. E. Macdonald, J. Ulstrup, and M. Elliott. Single-molecule mapping of long-range electron transport for a cytochrome b(562) variant. *Nano Lett.*, 11:176, 2011.

[46] N. J. Tao. Probing Potential-Tuned Resonant Tunneling through Redox Molecules with Scanning Tunneling Microscopy. *Phys. Rev. Lett.*, 76:4066, 1996.

[47] J. M. Artes, I. Diez-Perez, and P. Gorostiza. Transistor-like behavior of single metalloprotein junctions. *Nano Lett.*, 12:2679, 2012.

[48] A. M. Kuznetsov and J. Ulstrup. Single-molecule electron tunnelling through multiple redox levels with environmental relaxation. *J. Electroanal. Chem.*, 564: 209, 2004.

[49] I. V. Pobelov, Z. Li, and T. Wandlowski. Electrolyte gating in redox-active tunneling junctions–an electrochemical STM approach. *J. Am. Chem. Soc.*, 130:16045, 2008.

[50] Sara H Mejías, Begoña Sot, Raul Guantes, and Aitziber L Cortajarena. Controlled nanometric fibers of self-assembled designed protein scaffolds. *Nanoscale*, 6(19):10982–10988, 2014.

[51] J. Jortner, M. Bixon, T. Langenbacher, and M. E. Michel-Beyerle. Charge transfer and transport in DNA. *Proc. Natl. Acad. Sci. U. S. A.*, 95:12759, 1998.

[52] Sara H Mejias, Antonio Aires, Pierre Couleaud, and Aitziber L Cortajarena. Designed repeat proteins as building blocks for nanofabrication. *Protein-based Engineered Nanostructures*, pages 61–81, 2016.

[53] Ewan RG Main, Yong Xiong, Melanie J Cocco, Luca D'Andrea, and Lynne Regan. Design of stable $\alpha$-helical arrays from an idealized TPR motif. *Structure*, 11(5):497–508, 2003.

[54] Tommi Kajander, Aitziber L Cortajarena, Simon Mochrie, and Lynne Regan. Structure and stability of designed TPR protein superhelices: unusual crystal packing and implications for natural TPR proteins. *Acta Crystallographica Section D: Biological Crystallography*, 63(7):800–811, 2007.

[55] Jie Liu, Qi Zheng, Yiqun Deng, Neville R Kallenbach, and Min Lu. Conformational transition between four and five-stranded phenylalanine zippers determined by a local packing interaction. *Journal of Molecular Biology*, 361(1): 168–179, 2006.

[56] Jie Liu, Wei Yong, Yiqun Deng, Neville R Kallenbach, and Min Lu. Atomic structure of a tryptophan-zipper pentamer. *Proceedings of the National Academy of Sciences*, 101(46):16156–16161, 2004.

[57] S. Kamtekar, A. J. Berman, J. Wang, J. M. Lazaro, M. de Vega, L. Blanco, M. Salas, and T. A. Steitz. Insights into Strand Displacement and Processivity from the Crystal Structure of the Protein-Primed DNA Polymerase of Bacteriophage Phi29. *Mol. Cell*, 16:609, 2004.

[58] L. Blanco, A. Bernad, J. M. Lazaro, G. Martin, C. Garmendia, and M. Salas. Highly Efficient DNA Synthesis by the Phage Phi 29 DNA Polymerase: Symmetrical Mode of DNA Replication. *J. Biol. Chem.*, 264:8935, 1989.

[59] J. A. Esteban, M. Salas, and L. Blanco. Fidelity of Phi 29 DNA Polymerase. Comparison Between Protein-Primed Initiation and DNA Polymerization. *J. Biol. Chem.*, 268:2719, 1993.

[60] Medhanie Tesfay Gebrekidan, Christian Knipfer, Florian Stelzle, Juergen Popp, Stefan Will, and Andreas Braeuer. A shifted-excitation Raman difference spectroscopy (SERDS) evaluation strategy for the efficient isolation of Raman spectra from extreme fluorescence interference. *Journal of Raman Spectroscopy*, 47 (2):198–209, 2016.

[61] Yan Zhu, Shaghayegh Gharghabi, Diego Furtado Silva, Hoang Anh Dau, Chin-Chia Michael Yeh, Nader Shakibay Senobari, Abdulaziz Almaslukh, Kaveh Kamgar, Zachary Zimmerman, Gareth Funning, Abdullah Mueen, and Eamonn Keogh. The Swiss army knife of time series data mining: ten useful things you can do with the matrix profile and ten lines of code. *Data Mining and Knowledge Discovery*, 34:949–979, 2020.

[62] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[63] Zhenzhen Cheng, Lijun Qi, Yifan Cheng, Yalei Wu, and Hao Zhang. Interlacing Orchard Canopy Separation and Assessment using UAV Images. *Remote Sensing*, 12(5), 2020.

[64] Sibel Ebru Yalcin and Nikhil S. Malvankar. The blind men and the filament: Understanding structures and functions of microbial nanowires. *Current Opinion in Chemical Biology*, 59:193–201, 2020.

[65] Nicole L. Ing, Mohamed Y. El-Naggar, and Allon I. Hochbaum. Going the Distance: Long-Range Conductivity in Protein and Peptide Bioelectronic Materials. *The Journal of Physical Chemistry B*, 122(46):10403–10423, 2018.

[66] Tracy Q. Ha, Inco J. Planje, Jhanelle R.G. White, Albert C. Aragonès, and Ismael Díez-Pérez. Charge transport at the protein–electrode interface in the emerging field of BioMolecular Electronics. *Current Opinion in Electrochemistry*, 28:100734, 2021.

[67] Dmitry V. Matyushov. Solvent reorganization energy of electron-transfer reactions in polar solvents. *The Journal of Chemical Physics*, 120(16):7532–7556, 2004.

[68] D J Kushner, Alison Baker, and T G Dunstall. Pharmacological uses and perspectives of heavy water and deuterated compounds. *Canadian Journal of Physiology and Pharmacology*, 77(2):79–88, 1999.

[69] Patrizia Cioni and Giovanni B. Strambini. Effect of Heavy Water on Protein Flexibility. *Biophysical Journal*, 82(6):3246–3253, 2002.

[70] Judith P. Klinman and Amnon Kohen. Hydrogen Tunneling Links Protein Dynamics to Enzyme Catalysis. *Annual Review of Biochemistry*, 82(1):471–496, 2013.

[71] Michael J Weaver and Scott M Nettles. Solvent isotope effects upon the thermodynamics of some transition-metal redox couples in aqueous media. *Inorganic Chemistry*, 19(6):1641–1646, 1980.

[72] Ole Farver, Jingdong Zhang, Qijin Chi, Israel Pecht, and Jens Ulstrup. Deuterium isotope effect on the intramolecular electron transfer in *Pseudomonas aeruginosa* azurin. *Proceedings of the National Academy of Sciences*, 98(8): 4426–4430, 2001.

[73] Daniel H. Murgida and Peter Hildebrandt. Proton-Coupled Electron Transfer of Cytochrome c. *Journal of the American Chemical Society*, 123(17):4062–4068, 2001.

[74] Martin Byrdin, Valérie Sartor, André P.M. Eker, Marten H. Vos, Corinne Aubert, Klaus Brettel, and Paul Mathis. Intraprotein electron transfer and proton dynamics during photoactivation of DNA photolyase from E. coli: review and new insights from an "inverse" deuterium isotope effect. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1655:64–70, 2004.

[75] Setare Mostajabi Sarhangi and Dmitry V. Matyushov. Effect of Water Deuteration on Protein Electron Transfer. *The Journal of Physical Chemistry Letters*, 14(3):723–729, 2023.

[76] Juan M. Artés, Ismael Díez-Pérez, Fausto Sanz, and Pau Gorostiza. Direct Measurement of Electron Transfer Distance Decay Constants of Single Redox Proteins by Electrochemical Tunneling Spectroscopy. *ACS Nano*, 5(3):2060–2066, 2011.

[77] Marta P. Ruiz, Albert C. Aragonès, Nuria Camarero, J. G. Vilhena, Maria Ortega, Linda A. Zotti, Rubén Pérez, Juan Carlos Cuevas, Pau Gorostiza, and Ismael Díez-Pérez. Bioengineering a Single-Protein Junction. *Journal of the American Chemical Society*, 139(43):15337–15346, 2017.

[78] Anna Lagunas, Alejandra Guerra-Castellano, Alba Nin-Hill, Irene Díaz-Moreno, Miguel A De la Rosa, Josep Samitier, Carme Rovira, and Pau Gorostiza. Long distance electron transfer through the aqueous solution between redox partner proteins. *Nature Communications*, 9(1):5157, 2018.

[79] Peter J. Dahl, Sophia M. Yi, Yangqi Gu, Atanu Acharya, Catharine Shipps, Jens Neu, J. Patrick O'Brien, Uriel N. Morzan, Subhajyoti Chaudhuri, Matthew J. Guberman-Pfeffer, Dennis Vu, Sibel Ebru Yalcin, Victor S. Batista, and Nikhil S. Malvankar. A 300-fold conductivity increase in microbial cytochrome nanowires due to temperature-induced restructuring of hydrogen bonding networks. *Science Advances*, 8(19):eabm7193, 2022.

[80] Anthony Harriman. Further comments on the redox potentials of tryptophan and tyrosine. *Journal of Physical Chemistry*, 91(24):6102–6104, 1987.

[81] Michael Tuchband, Jin He, Shuo Huang, and Stuart Lindsay. Insulated gold scanning tunneling microscopy probes for recognition tunneling in an aqueous environment. *Review of Scientific Instruments*, 83(1):015102, 2012.

[82] Sara H Mejias, Pierre Couleaud, Santiago Casado, Daniel Granados, Miguel Angel Garcia, Jose M Abad, and Aitziber L Cortajarena. Assembly of designed protein scaffolds into monolayers for nanoparticle patterning. *Colloids and Surfaces B: Biointerfaces*, 141:93–101, 2016.

[83] Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7):1731–1737, 2003.

[84] Shuai Chang, Jin He, Peiming Zhang, Brett Gyarfas, and Stuart Lindsay. Gap Distance and Interactions in a Molecular Tunnel Junction. *Journal of the American Chemical Society*, 133(36):14267–14269, 2011.

[85] Edward L. Mertz and Lev I. Krishtalik. Low dielectric response in enzyme active site. *Proceedings of the National Academy of Sciences*, 97(5):2081–2086, 2000.

[86] Yoni Eshel, Uri Peskin, and Nadav Amdursky. Coherence-assisted electron diffusion across the multi-heme protein-based bacterial nanowire. *Nanotechnology*, 31(31):314002, 2020.

[87] Jake VanderPlas. *Python data science handbook: Essential tools for working with data*. O'Reilly Media, Inc., 2016.

[88] Siddharth Krishnan, Aleksei Aksimentiev, Stuart Lindsay, and Dmitry Matyushov. Long-range Hopping Conductivity in Proteins. *bioRxiv*, pages 2022–10, 2022.

[89] Juliette TJ Lecomte and Gerd N La Mar. Proton NMR study of labile proton exchange in the heme cavity as a probe for the potential ligand entry channel in myoglobin. *Biochemistry*, 24(25):7388–7395, 1985.

[90] Daniel Mark Shapiro, Gunasheil Mandava, Sibel Ebru Yalcin, Pol Arranz-Gibert, Peter J Dahl, Catharine Shipps, Yangqi Gu, Vishok Srikanth, Aldo I Salazar-Morales, J Patrick O'Brien, et al. Protein nanowires with tunable functionality and programmable self-assembly using sequence-controlled synthesis. *Nature Communications*, 13(1):829, 2022.

[91] Zdenek Futera, Ichiro Ide, Ben Kayser, Kavita Garg, Xiuyun Jiang, Jessica H. van Wonderen, Julea N. Butt, Hisao Ishii, Israel Pecht, Mordechai Sheves, David Cahen, and Jochen Blumberger. Coherent Electron Transport across a 3 nm Bioelectronic Junction Made of Multi-Heme Proteins. *The Journal of Physical Chemistry Letters*, 11(22):9766–9774, 2020.

[92] Masateru Taniguchi, Shohei Minami, Chikako Ono, Rina Hamajima, Ayumi Morimura, Shigeto Hamaguchi, Yukihiro Akeda, Yuta Kanai, Takeshi Kobayashi, Wataru Kamitani, et al. Combining machine learning and nanopore construction creates an artificial intelligence nanopore for coronavirus detection. *Nature Communications*, 12(1):3726, 2021.

[93] Akihide Arima, Makusu Tsutsui, Takashi Washio, Yoshinobu Baba, and Tomoji Kawai. Solid-state nanopore platform integrated with machine learning for digital diagnosis of virus infection. *Analytical Chemistry*, 93(1):215–227, 2020.

[94] Mohamed Soliman Halawa, Rebeca P. Díaz Redondo, and Ana Fernández Vilas. Kpis-based clustering and visualization of hpc jobs: A feature reduction approach. *IEEE Access*, 9:25522–25543, 2021.

[95] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*, 307:72–77, 2018.

[96] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[97] Nikhil Ketkar and Nikhil Ketkar. Stochastic gradient descent. *Deep learning with Python: A hands-on introduction*, pages 113–132, 2017.

[98] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

[99] Bintian Zhang, Hanqing Deng, Sohini Mukherjee, Weisi Song, Xu Wang, and Stuart Lindsay. Engineering an Enzyme for Direct Electrical Monitoring of Activity. *ACS Nano*, 14(2):1360–1368, 2020.

[100] L. Blanco and M. Salas. Relating Structure to Function in Phi29 DNA Polymerase. *J. Biol. Chem.*, 271:8509, 1996.

[101] J. Korlach, A. Bibillo, J. Wegener, P. Peluso, T. T. Pham, I. Park, S. Clark, G. A. Otto, and S. W. Turner. Long, Processive Enzymatic DNA Synthesis Using 100% Dye-Labeled Terminal Phosphate-Linked Nucleotides. *Nucleosides, Nucleotides Nucleic Acids*, 27:1072, 2008.

[102] T. A. Steitz. A Mechanism for All Polymerases. *Nature*, 391:231, 1998.

[103] M. Fairhead and M. Howarth. Site-Specific Biotinylation of Purified Proteins Using BirA. *Methods Mol. Biol.*, 1266:171, 2015.

[104] A. J. Berman, S. Kamtekar, J. L. Goodman, J. M. Lazaro, M. de Vega, L. Blanco, M. Salas, and T. A. Steitz. Structures of Phi29 DNA Polymerase Complexed with Substrate: the Mechanism of Translocation in B-family Polymerases. *EMBO J.*, 26:3494, 2007.

[105] P. Ghisellini, M. Caiazzo, A. Alessandrini, R. Eggenhoffner, M. Vassalli, and P. Facci. Direct Electrical Control of IgG Conformation and Functional Activity at Surfaces. *Sci. Rep.*, 6:37779, 2016.

[106] D. Beckett, E. Kovaleva, and P. J. Schatz. A Minimal Peptide Substrate in Biotin Holoenzyme Synthetase-Catalyzed Biotinylation. *Protein Sci.*, 8:921, 1999.

[107] W. A. Beard, D. D. Shock, V. K. Batra, L. C. Pedersen, and S. H. Wilson. DNA Polymerase Beta Substrate Specificity: Side Chain Modulation of the "A - Rule". *J. Biol. Chem.*, 284:31680, 2009.

[108] J. A. Morin, F. J. Cao, J. M. Lazaro, J. R. Arias-Gonzalez, J. M. Valpuesta, J. L. Carrascosa, M. Salas, and B. Ibarra. Mechano-Chemical Kinetics of DNA Replication: Identification of the Translocation Step of a Replicative DNA Polymerase. *Nucleic Acids Res.*, 43:3643, 2015.

[109] J. M. Lazaro, L. Blanco, and M. Salas. Purification of Bacteriophage Phi 29 DNA Polymerase. *Methods Enzymol.*, 262:42, 1995.

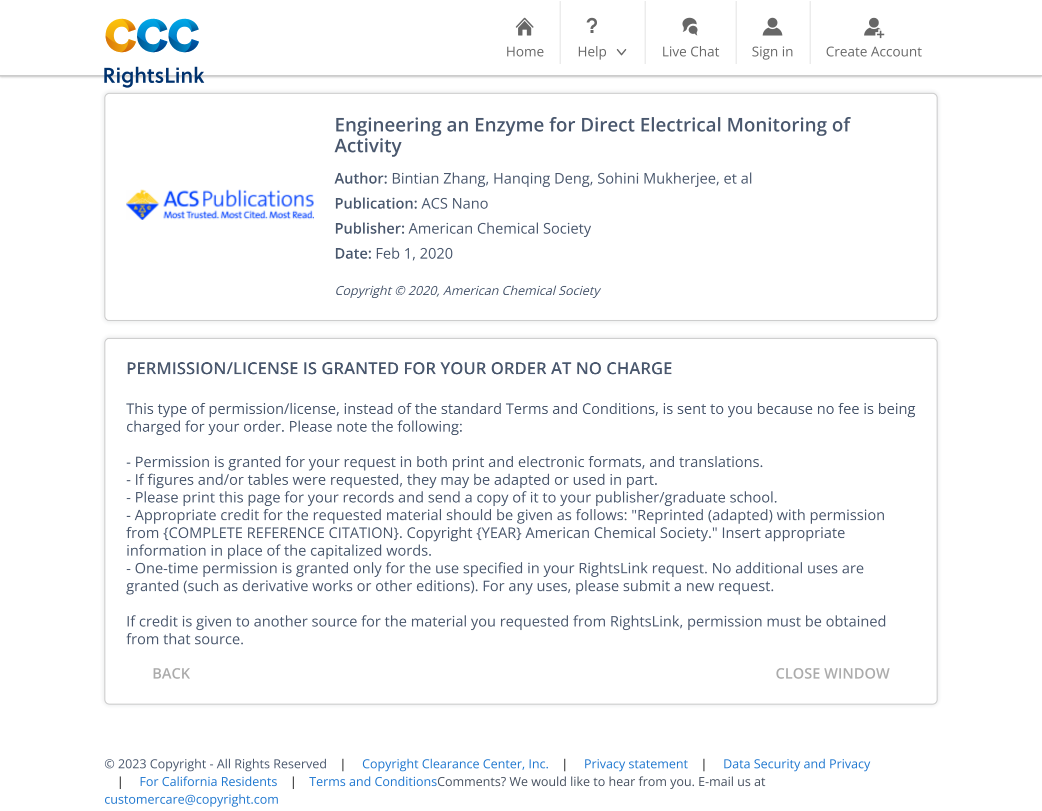
[110] J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, and J. Tan. Asymmetric Least Squares for Multiple Spectra Baseline Correction. *Anal. Chim. Acta*, 683:63, 2010.

[111] Andrew Van Benschoten, Austin Ouyang, Francisco Bischoff, and Tyler Marrs. MPA: a novel cross-language API for time series analysis. *Journal of Open Source Software*, 5(49):2179, 2020.

[112] P. Pang, B. A. Ashcroft, W. Song, P. Zhang, S. Biswas, Q. Qing, J. Yang, R. J. Nemanich, J. Bai, J. T. Smith, K. Reuter, V. S. K. Balagurusamy, Y. Astier, G. Stolovitzky, and S. Lindsay. Fixed-Gap Tunnel Junction for Reading DNA Nucleotides. *ACS Nano*, 8:11994, 2014.

# APPENDIX A

# PERMISSIONS TO USE COPYRIGHTED MATERIALS



Engineering an Enzyme for Direct Electrical Monitoring of Activity

**Author:** Bintian Zhang, Hanqing Deng, Sohini Mukherjee, et al

**Publication:** ACS Nano

**Publisher:** American Chemical Society

**Date:** Feb 1, 2020

*Copyright © 2020, American Chemical Society*

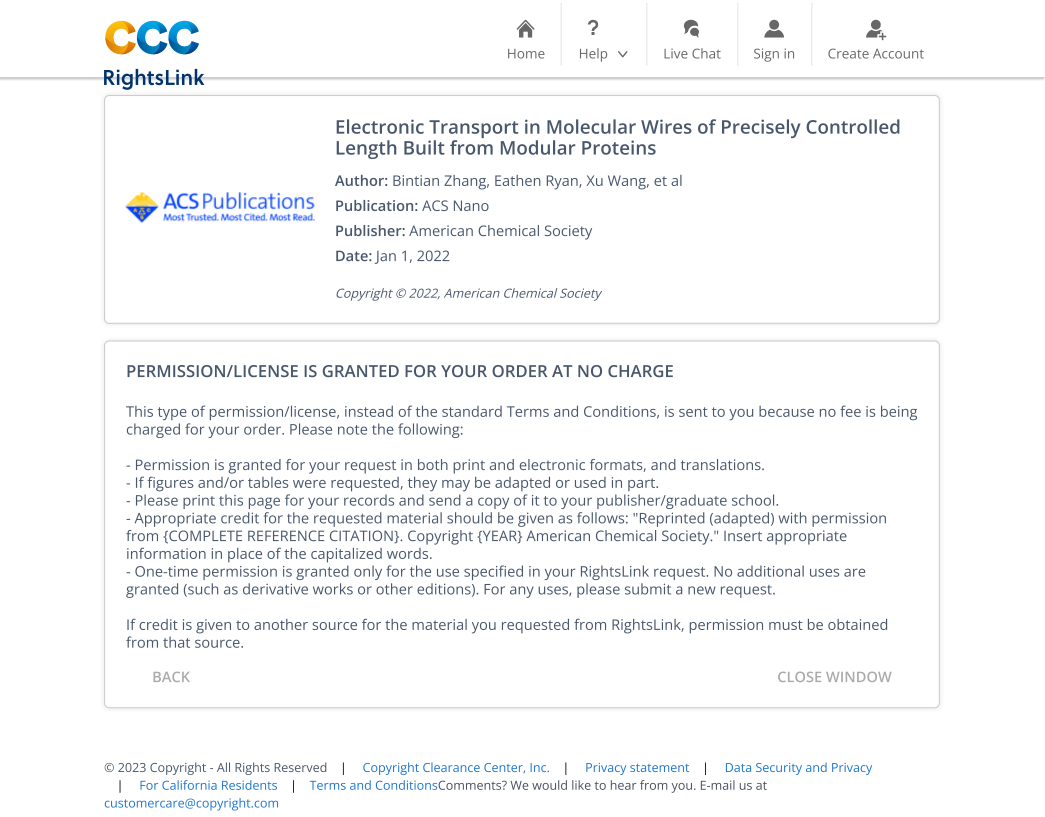
PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms and Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from {COMPLETE REFERENCE CITATION}. Copyright {YEAR} American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your RightsLink request. No additional uses are granted (such as derivative works or other editions). For any uses, please submit a new request.

If credit is given to another source for the material you requested from RightsLink, permission must be obtained from that source.

BACK CLOSE WINDOW



## Electronic Transport in Molecular Wires of Precisely Controlled Length Built from Modular Proteins

**Author:** Bintian Zhang, Eathen Ryan, Xu Wang, et al

**Publication:** ACS Nano

**Publisher:** American Chemical Society

**Date:** Jan 1, 2022

*Copyright © 2022, American Chemical Society*

### PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms and Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from {COMPLETE REFERENCE CITATION}. Copyright {YEAR} American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your RightsLink request. No additional uses are granted (such as derivative works or other editions). For any uses, please submit a new request.

If credit is given to another source for the material you requested from RightsLink, permission must be obtained from that source.

# APPENDIX B

# STATEMENT OF CO-AUTHOR PERMISSIONS

All the co-authors and collaborators have granted their consent for sharing of data in Chapter 2, 3 and 4.

# APPENDIX C

# STATEMENT OF CODE AVAILABILITY

The Python-based scripts used in Chapter 2, 3 and 4 are given in `https://github.com/sohini0512/Single-Molecule-Protein-Conductance-Measurements`.