UNIVERSITY OF CALIFORNIA,
IRVINE

Single-Molecule Studies of Biomolecules as Molecular Machines

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Physics

by

Calvin James Lau

Dissertation Committee:
Professor Philip G. Collins, Chair
Professor Zuzanna S. Siwy
Assistant Professor Albert Siryaporn

2020

DEDICATION

To

my father

who, though now reunited with our Heavenly Father,

gave me a love for knowledge and taught me the importance of careful and critical thinking,

and who exemplified a perspective that unified both science and faith in the pursuit of scientific knowledge and an understanding of our world.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

As this chapter of my life comes to a close, I would like to take the time to thank the many people who have blessed my life and without whom this work would not be possible. I have received so much patience, guidance, encouragement, and support from so many people, and I am so grateful for their presence in my life through the past six years.

Professor Philip Collins has been an incredible advisor and mentor, providing stability and structure in the research laboratory and also much professional guidance. I am grateful for his vision for my success, his patience with me, and his willingness to challenge my mindset for growth. Over the years, I have grown to appreciate how intentional he is in coaching each of his students to grow in their critical thinking abilities, communication and presentation skills, and productivity.

Professor Gregory Weiss has been an enthusiastic and supportive presence throughout the years of working together. I appreciate his patience in explaining the many biology-related concepts that were new when I first started, as well as his words of support and encouragement during the lows of research and life.

I want to thank Dr. Patrick Sims and Dr. Max Akhterov, who undertook the work of training me when I first joined the Collins research group. I appreciate their patience and taking the time to answer the many questions that I asked.

I am grateful for, and indebted to, the two postdoctoral researchers who partnered with me and shared the burden for the measurement sections of this work: Dr. Wonbae Lee for the antibody binding research and Dr. Narendra Kumar for the φ29 DNA polymerase research. I am especially grateful for their patience and perseverance, both in preparing devices and staring at the computer screen for hours on end to watch for protein activity. Certainly, without their presence and hard work, these results would not exist today.

I am also grateful for the hard work of many members of the Weiss research group, including Dr. Mark Richardson, Dr. Sudipta Majumdar, Kristin Gabriel, Rebekah Dyer, Jessica Fong, Joshua Kim, and others, who labored hard to get the proteins expressed, purified, and delivered, often on short notice, and who also graciously provided all the substrates used in this work. I am particularly indebted to the wonderful Kristin Gabriel, who oversaw the φ29 and Taq DNA polymerase expression and mutagenesis and who has been a very supportive and encouraging partner throughout.

Further appreciation goes to Arith Rajapakse, Davil Garcia, and Lauren Brooks, who fabricated the SWCNT-FET devices used throughout this work, as well as the former Collins lab members who built the laboratory infrastructure that I inherited. I also want to thank the other former and current Collins and Weiss lab members with whom I overlapped in tenure, including Dr. Deng Pan, Dr. Elliot Fuller, Dr. Tetyana Ignatova, Dr. Denys Marushchak, Jill Pestana, Mackenzie Turvey, Dr. Mina Baghgar, Jeffrey Taulbee, Dr. Mariam Iftikhar, Dr. Joshua Smith, and Dr. Kaitlin Pugliese, for providing both guidance for

navigating the lab as well as encouragement and support. In particular, I would like to thank Mackenzie Turvey for the many conversations, random stories, funny jokes, and general commiserating during our 4.5+ years of shared experience in the Collins lab.

I want to thank my parents for their love and support throughout the years, and especially my mom for cooking so much food and bringing it over when I didn't have time to cook for myself. Their love and prayers and unconditional support gave me strength to continue.

Additional thanks goes to the Lau cousins who preceded me in graduate studies at UCI, and for the advice, wisdom, and encouragement that they gave to me when I first started as a graduate student.

I also want to thank my housemates over the years, as well as my many friends and mentors in both Acts2Fellowship and International Graduate Student Ministry, for their prayers, their patience with me, and their emotional support. I am grateful for their kind words of encouragement and understanding, and their bearing with me through the exhaustion and late nights.

Finally, but most importantly, I would like to thank God, to whom I owe my entire existence and every gift and ability I possess. Without His gracious forgiveness and leading of my life, I would not be here today.

The work in Chapter 2 was made possible by a generous sponsorship from Autotelic Inc., while the work in Chapters 3-5 was funded by NIH grant R01-HG009188.

VITA

Calvin James Lau

2011-12	Undergraduate Student Researcher, University of California, Berkeley
2012	B.A. in Physics, University of California, Berkeley
2012-14	Junior Scientist, Porifera, Inc.
2014-15	Teaching Assistant, Department of Physics and Astronomy, University of California, Irvine
2015-20	Graduate Student Researcher, Department of Physics and Astronomy, University of California, Irvine
2019	M.S. in Physics and Astronomy, University of California, Irvine
2020	Ph.D. in Physics and Astronomy, University of California, Irvine

FIELD OF STUDY

Single-Molecule Biophysics

PUBLICATIONS

Richardson MB, Gabriel KN, Garcia J, Ashby S, Dyer R, Kim J, Lau C, Hong J, Le Tourneau RJ, Sen S, Narel D. Pyrocinchonimides Conjugate to Amine Groups on Proteins via Imide Transfer. Bioconjugate Chemistry. 2020 Apr 17.

ABSTRACT OF THE DISSERTATION

Single-Molecule Studies of Biomolecules as Molecular Machines

By

Calvin James Lau

Doctor of Philosophy in Physics

University of California, Irvine, 2020

Professor Philip G. Collins, Chair

Studies of single biomolecules provide information that is buried in an ensemble-based measurement, including the evolution of an individual biomolecule's behavior over time. Recent work showed that an electronic sensor composed of single-walled carbon nanotube field-effect transistors (SWCNT-FETs) can observe an individual biomolecule's conformational motions over time and obtain accurate measurements of catalytic rates for a variety of enzymes (1-3).

This dissertation expands the scope of transistor-based biosensing techniques through several strategies. The first strategy extends previous work by investigating similar enzymes, such as other DNA polymerases, in order to identify unique characteristics of each enzyme. A second strategy focuses on investigating the dynamics of biomolecular interactions that have not been previously studied by this technique, such as ligand-binding interactions. A third strategy makes refinements to the measurement or analysis techniques to uncover additional, subtler dynamics and other information that was previously hidden in the acquired signal.

The dissertation is organized into two main parts. The first part (Chapters 1-3) discusses new measurements performed using the SWCNT-FET technique. Chapter 1 provides a brief introduction to the SWCNT-FET biosensing technique and details the methods and materials used in the experiments described in the following two chapters.

Chapter 2 studies the behavior of a weakly-interacting antibody-antigen system: antibody 3C6 in the presence of the antigen paclitaxel. SWCNT-FET recordings of antigen-antibody binding exhibited two conductance states corresponding to bound and unbound configurations, like the two-level dynamics previously recorded from enzymatic catalysis. The SWCNT-FET signal correlated with antigen concentration, remaining relatively static at concentrations far from the value of the dissociation constant K_D and fluctuating most actively near K_D. Analysis of the distribution of single-molecule bound and unbound times determined a value of K_D = 30 nM, a binding rate $k_{off} = 10^4\ s^{-1}$, and a Hill coefficient of binding cooperativity of 1.8. Chapter 2 also compares antibody-antigen dynamics recorded in single-molecule, few-molecule, and many-molecule regimes of biofunctionalization.

Chapter 3 extends previous work on DNA polymerases by investigating an alternate polymerase, φ29 DNA polymerase, and characterizes its conformational motions and catalytic efficiency. Chapter 3 finds that the catalytic efficiency of φ29 DNA polymerase depends on the template composition. The enzyme continuously processed heteropolymer ssDNA templates and homopolymer templates containing thymine and cytosine at rates of ~50 s^{-1} for 3-5 mins, but exhibited only 1-2 s bursts of conformational motion among 60 s of

pauses when processing homopolymer templates containing adenine and guanine. Single-molecule recordings of the latter two templates showed the ability of the SWCNT-FET to measure enzyme motion when the enzyme's activity in time was less than 2%, in contrast with ensemble-based observations which did not detect catalysis at such low activity. In addition, detailed analyses of the open and closed conformations of φ29 DNA polymerase suggested the presence of multiple operating modes during the catalytic cycle, including a closed conformation whose duration was 8 times longer than the typical catalytic event.

The second part of the dissertation (Chapters 4-5) expands the analysis toolkit with new methods. Chapter 4 introduces a wavelet-based denoising scheme and describes the optimization and application of the scheme to SWCNT-FET sensor signals. This chapter demonstrates the effectiveness of wavelet denoising in removing both low- and high-frequency noise from the SWCNT-FET sensor output and describes the artifacts introduced by the denoising process. The wavelet denoising scheme is also compared to other digital denoising methods.

Finally, Chapter 5 describes an automated analysis procedure for identifying conformational events and characterizing each event using a set of features. SWCNT-FET measurements from two variants of Taq DNA polymerase are compared to highlight features that are correlated to enzyme behavior, correlated to experimental noise, or completely uncorrelated. In addition, a preliminary analysis using principal component analysis (PCA) serves as an example of machine learning techniques that could be used in the future.

CHAPTER 1

Experimental Methods

1.1 Introduction

Single-molecule techniques provide a way to observe the variations in biomolecule dynamics and interactions that are otherwise obscured by ensemble averaging. The techniques used to study individual biomolecules and their characteristics include: optical tweezers (4-11), fluorescence (including single-molecule fluorescence resonance energy transfer (smFRET) (4, 8, 12-15) and protein induced fluorescence enhancement (PIFE) (16, 17)), probe microscopy and tunneling sensing (including atomic force microscopy (AFM) (4, 5), scanning tunneling microscopy (STM) (18, 19), and recognition tunneling (20-22)), nanopore-based sensing (23, 24), and transistor-based sensing (1, 25-28). Each technique can be loosely categorized by the type of information obtained: structure (STM, recognition tunneling), force characteristics (optical tweezers, AFM), conformational dynamics (smFRET, PIFE, transistor-based sensing), and enzyme processive dynamics (AFM, optical tweezers, and nanopore- and transistor-based sensing).

Previous work demonstrated a transistor-based method for monitoring the conformational dynamics of single enzymes by tethering the enzyme to a single-walled carbon nanotube field-effect transistor (SWCNT-FET) and measuring changes in electrical conductance (1-3). Studies of T4 lysozyme (1) and DNA polymerase I (2) showed that conformational changes in the enzyme are transduced into fluctuations in the transistor conductance via local field-

effect gating from charged residues near the tether location (25). These experiments produced results similar to those of previous studies of the same enzymes using different methods, confirming that this technique records signals correlated to enzyme activity.

All the techniques that probe conformational or processive dynamics require attaching the biomolecule to a tether, to attach the molecule to a fluorophore label (smFRET, PIFE) or to an optical bead (optical tweezers) or to a nanopore or transistor sensor. Each technique has tradeoffs. Fluorescence techniques can measure individual molecules either at high temporal resolutions (under 1 ns) at short measurement durations (less than 1 s) or multiple molecules simultaneously at low temporal resolutions (~100 μs) for up to several minutes. Optical tweezers can observe individual molecules for long periods of time (>50 min) but suffer from low temporal resolution (>100 μs) and are limited to one molecule at a time. Transistor-based techniques, including SWCNT-FET sensing, have both high time resolution (~1 μs), and long continuous measurement durations (>60 min), but the signal-to-noise ratio is limited by the arrangement of charged residues near the tether position. In addition, transistor-based techniques are the easiest to implement for commercial-level molecular sensing and DNA sequencing, being highly scalable (to hundreds or thousands of individual sensing elements) and easily integrated with digital electronics.

The SWCNT-FET sensor records the history of an individual biomolecule over time. The combination of long-duration measurements and high temporal resolution gives the SWCNT-FET the ability to directly observe the interactions or correlations between long-duration and short-duration behaviors. Examples include: changes in the distributions of

conformation dwell times over multiple minutes, and transient events that stop a biomolecule's normal activity, causing it to pause for multiple minutes. In addition, the SWCNT-FET technique can observe sporadic behaviors that are too rare to appear in any ensemble-based measurement, such as an enzyme performing catalysis that is statistically or energetically unfavored.

The remainder of this chapter describes the experimental protocols for the SWCNT-FET measurement technique, including device fabrication, biomolecule conjugation, measurement, and data analysis.

1.2 SWCNT Synthesis and Device Fabrication

The SWCNT-FET devices are fabricated by depositing metal electrodes by photolithography and electron beam evaporation onto CNTs grown across 4" SiO2/Si wafers using chemical vapor deposition (CVD). During the CVD process, the CNTs grow as the carbon feedstock, methane, decomposes into amorphous carbon and nucleates around randomly-distributed Fe-Mo catalyst particles. For some experiments, a layer of alumina (Al_2O_3) is laid down on top of the entire wafer using atomic layer deposition (ALD) to protect the CNTs from contamination during subsequent processing. A bilayer photoresist (Shipley 1808 on top of LOR-A1 from MicroChem) is deposited onto the wafer by spin coating, and the electrode geometry is transferred to the resist using a patterned mask and UV exposure. Any alumina in the electrode pattern is removed with Transetch N (Transene), and the electrode metal is deposited using evaporation: first a titanium sticking layer, then ~10-20 nm of platinum.

After the photoresist and excess metal are removed with Remover PG (Microchem), the SWCNT-FET devices are completed, composed of a single SWCNT covered by two electrodes.

Each device on the wafer is probed to electrically characterize all the CNT connections between electrodes, and then the wafer is diced into individual chips of ~1 cm by ~1 cm. Individual SWCNT-FET devices exhibiting an on-resistance of 100-500 kΩ and a current on:off ratio of >100:1 are selected. Each chip containing a selected device is cleaned with Remover PG at 60°C to remove as much leftover photoresist as possible. A ~300 nm layer of PMMA (A3 PMMA from MicroChem) is deposited onto the chip surface by spin coating to electrically and mechanically isolate the metal electrodes from exposure to the test solutions and molecules. A 1 μm-wide trench is created in the PMMA in the gap between electrodes, patterned by electron beam lithography and opened using resist developer (a mixture of isopropyl alcohol and methyl isobutyl ketone) to expose only the SWCNT sidewall. For those devices covered by an ALD layer of alumina, the remaining PMMA layer is baked at 194°C to harden the PMMA, then the chip is treated with diluted Transetch N (Transene) to etch away the ALD directly in the trench, exposing the CNT sidewall.

1.3 Electrical Measurements

The electrical measurements described throughout this work were conducted with two separate experimental setups. The measurements of the antibody-paclitaxel interaction (Chapter 2) and the comformational behavior of φ29 DNA polymerase (Chapter 3) were conducted on the probe station, for which the device is simply covered by a ~1-2 μL drop of

buffer solution that is open to air. Measurements on the probe station are conducted at room temperature (~22°C). The measurements of Taq DNA polymerase (used as examples in Chapters 4 and 5) were conducted in a flow-cell setup, in which the device is sealed in a ~5 mm long microfluidic channel by a PDMS gasket that limits the channel cross-section to ~100 μm by ~100 μm. The smaller liquid volume in the flow cell (~0.5 μL) reduces the capacitance of the liquid, reducing the effective time constant $\tau = RC$ of the experimental setup and allowing stable measurements at higher bandwidth. The fluid temperature of the flow-cell experiment can be controlled from 22°C to 94°C.

All electrical measurements are performed with the biomolecule and exposed SWCNT sidewall (the active portion of the device) submerged in solution. The solution potential is maintained by platinum reference and counter electrodes, with the voltage between the reference electrode and the SWCNT (hereafter called liquid-gate voltage, or V_g) maintained by a Keithley 2400 SourceMeter. The back gate (back surface of the device) is maintained at ground (0 V), and a voltage (V_{sd}) is applied between the source and drain electrodes on the chip surface. The resulting electric current passing through the SWCNT-FET, hereafter called the $I(t)$, is amplified and recorded.

In the probe station, the $I(t)$ is amplified by a Keithley 428 current amplifier set to a gain of 10^8 V/A (with a 10%-90% rise time of 40 μs for an effective bandwidth of 25 kHz), and the resulting signal is acquired by a National Instruments data acquisition card (PCI-6281 or PCIe-6361) at 100 kHz. In the flow cell, the $I(t)$ is amplified with a Femto DLPCA-200 or

DHPCA-100 current amplifier at a bandwidth between 200 MHz and 1.8 GHz and acquired with a National Instruments PCIe-6361 DAQ card at 1 MHz.

The resulting SWCNT-FET devices are electrically characterized in buffer solution as a control. Generally, carbon nanotube transistors have inherent 1/f (pink) noise (29), and sometimes also exhibit random telegraph switching (RTS) noise (30, 31) without any additional molecules attached. Such RTS noise is caused by defects in the SWCNT, charge traps in the underlying SiO_2 substrate on which the CNTs are fabricated, or impurities at the CNT-metal contact. Unfortunately, RTS noise can appear similar to switching signal generated by biomolecule activity. Thus, any SWCNT-FET/biomolecule complex that exhibited RTS before biomolecule conjugation was excluded from further measurement.

First, time-averaged current versus liquid gate potential (henceforth termed I-V_g) measurements are taken, with -0.4V < V_g < +0.6V, resulting in I-V_g curves with each point being an average of 0.5 s of data at 5 kHz, to determine the threshold voltage (V_T) for the SWCNT-FET device. Such I-V_g measurements correspond to those performed by other CNT and graphene biosensor experiments (27, 28, 32). Then, the V_g is held constant and $I(t)$ observed for a minimum of 5 min at various manually-chosen V_g values to confirm the absence of RTS before biomolecule conjugation.

After biomolecule conjugation (detailed in Section 1.4), $I(t)$ measurements of the biomolecule activity are acquired for various constant V_g values, with each measurement having a minimum duration of 5 min (and often extending beyond 10 min).

1.4 Linker and Biomolecule Conjugation

N-(1-pyrenyl)maleimide (Sigma-Aldrich), hereafter called pyrene-maleimide, is used as a linker molecule to anchor the 3C6 antibody to the SWCNT sidewall. The pyrene group, consisting of four fused benzene rings, non-covalently bonds to the SWCNT sidewall by π-π stacking, while the maleimide group forms stable thioether bonds with the free thiol of a cysteine on the antibody molecule (33, 34). Pyrene-maleimide is insoluble in water, and barely soluble in ethanol, but readily dissolves in non-polar solvents such as dimethyl sulfoxide (DMSO) and dichloromethane (DCM). Tests showed that ethanol, DMSO, and DCM are all suitable solvents for transferring pyrene-maleimide to the CNT sidewall, although DMSO and DCM dissolve the PMMA passivation layer and are thus incompatible with any experiments requiring PMMA passivation. If either DMSO or DCM are used as the solvent, any PMMA on the chip surface must be removed using Remover PG before linker attachment.

Mixtures of pyrene and pyrene-maleimide in solution provide a way to control the attachment density of biomolecules on the CNT. Both pyrene and pyrene-maleimide molecules adhere to the CNT sidewall, but only the pyrene-maleimide bonds with the biomolecule. A schematic of the SWCNT-FET device with both pyrene-maleimide and pyrene attached to the CNT sidewall is shown in Figure 1.1. A pyrene-maleimide:pyrene ratio of 1:10,000 results in an attachment density of 2-4 biomolecules for every μm of CNT sidewall, while a smaller ratio lowers the density and a higher ratio increases the density.

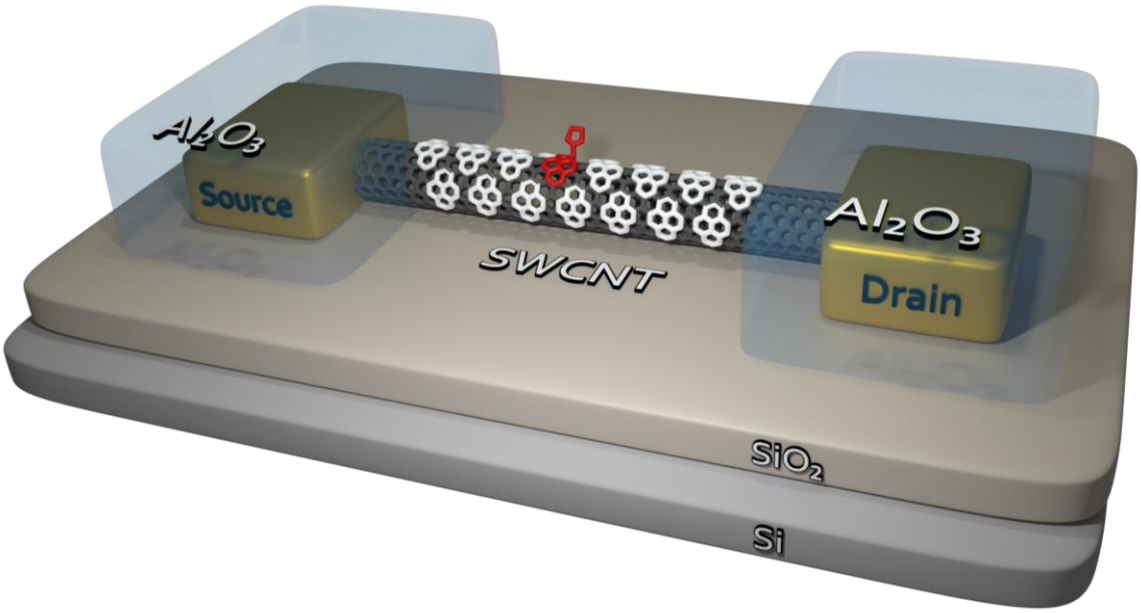

Figure 1.1: Schematic of the SWCNT-FET, with the source and drain electrodes (gold) passivated by Al_2O_3 (light blue), having both pyrene-maleimide (red) and pyrene (white) conjugated to the sidewall of the CNT (dark gray). The entire device rests on the surface of a silicon wafer (light gray) passivated by an oxide layer (SiO_2, in tan).

The pyrene-maleimide linker molecule is conjugated to the SWCNT-FET device by immersing the entire chip (containing the SWCNT-FET device) in a pyrene-maleimide solution for 2 min, then rinsing for 30 s to remove excess linker. The rinsing protocol depends on the solution used to dissolve the pyrene-maleimide. When DMSO is the solvent, the chips are rinsed with acetonitrile, then isopropyl alcohol (IPA), and then deionized (DI) water, each for 10 s. When ethanol is the solvent, the chips are rinsed with a solution of 0.1% Tween-20 in ethanol, then with DI water, each for 10 s.

Once the pyrene-maleimide is conjugated to the sidewall of the SWCNT-FET device, the biomolecules to be tested are conjugated to the pyrene-maleimide molecules in solution. The biomolecule sample is diluted from its storage solution to ~4 nM in attachment buffer

(although nominal concentrations of up to 400 nM may be used to counteract biomolecule aggregation as the sample ages). The ~4 nM biomolecule solution is pipetted onto the surface of the chip and left at room temperature. After 5 min of incubation, the excess solution is rinsed off, and the chip is stored in buffer until taken out immediately before electrical measurements.

1.5 Signal Processing and Analysis

The raw $I(t)$ is generally composed of short, sharp spikes superimposed on a drifting baseline whose mean $<I(t)>$ constantly changes in time, as shown in Figure 1.2A. This makes finding discrete states in the signal difficult, because the low-frequency drift changes the position of each state from moment to moment. To make the baseline position consistent, the signal is first detrended (Figure 1.2B) by removing low-frequencies from the signal: $\Delta I(t) = I(t) - < I(t) >$, where $<I(t)>$ is calculated with a low-pass filter, which brings the baseline state to 0. Further denoising (Figure 1.2C) removes some high-frequency noise components to reduce the width of the noise band in the baseline and to assist in finding the boundaries between states (shown as the red line in Figure 1.2D). Further discussion of $I(t)$ signal denoising and a comparison of different denoising methods is given in Chapter 4.

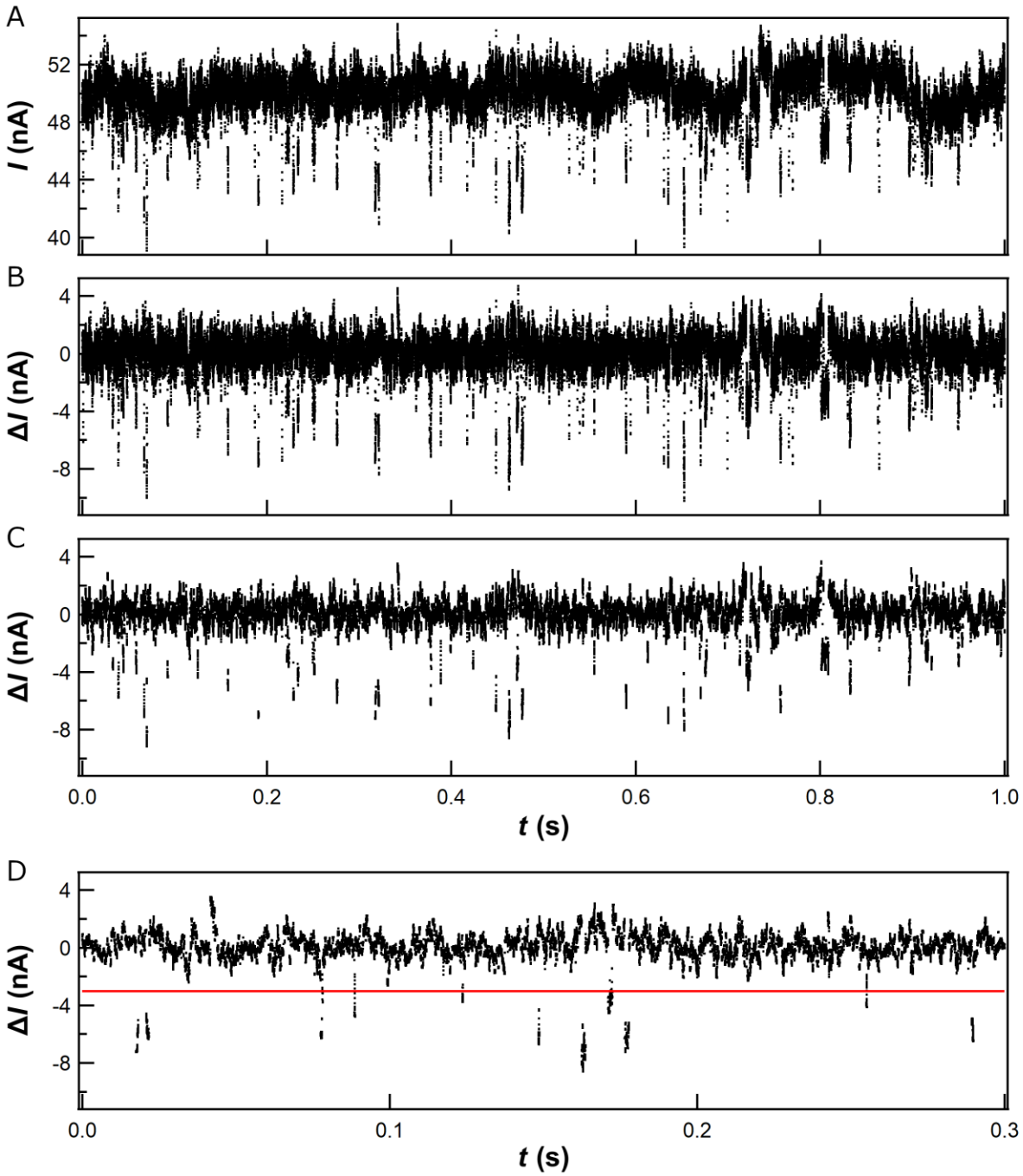

Figure 1.2: Examples of (A) raw $I(t)$, (B) detrended $\Delta I(t)$, and (C) detrended and denoised $\Delta I(t)$ from the SWCNT-FET sensor. In each case, the data points mostly reside in a fluctuating baseline, with occasional excursions away from the baseline. (D) A zoomed-in portion of the detrended and denoised signal, with a red line showing the position of the boundary between states. Points above this boundary are in the baseline state, and points below are in the excursion, or secondary, state.

Once the boundaries between states are established, each data point in the $\Delta I(t)$ is assigned to a state based on its position, and consecutive samples residing in the same state are counted to calculate the duration of each state instance. The threshold algorithm also ignores any events shorter than some minimum duration (which is dependent on the bandwidth of the current amplifier used during measurement) to reject spurious spike-like noise peaks that are too short to have passed through the current amplifier. Further discussion of defining the state boundaries and identifying events is given in Chapter 5. All the signal processing procedures were implemented in LabVIEW, and a brief description of the LabVIEW programs is contained in Appendix B.

When a biomolecule is attached to a SWCNT-FET biosensor, changes in the biomolecule's conformation appear as sharp transitions between discrete levels in the $I(t)$. One primary assumption for analyzing signals transduced from biomolecule motion is that there are a finite number (often three or less) of possible states present in the signal, with each state defined by a position, or amplitude away from the baseline. By virtue of being the most probable state, the baseline state is assumed to be the default, resting state of the biomolecule, which could be the catalytically inactive conformation for an enzyme or the equilibrium binding state of an antibody or receptor protein. Then, any deviations of the signal from the baseline to another state is an indication of biomolecule activity: a catalytic event, binding or unbinding event, or other conformational change.

Enzymes like lysozyme or DNA polymerase possess one or more catalytic cycles, through which the enzymes progress as they operate. Within a single cycle, an enzyme may transition

between two or more distinct conformational states, returning to the initial state when the cycle is complete. An enzyme spends the largest amount of time in the conformation before the slowest step of the catalytic cycle, and the state corresponding to that conformation is designated as the baseline, default state. For such enzymes, a conformational event is defined as a sequence of two transitions (a first transition from the baseline state to a different state, and a second transition back to the baseline state) which indicates that the enzyme has progressed through a portion of the catalytic cycle containing a change in conformation. Typically, the time spent in the non-baseline state is at least an order of magnitude smaller than the time spent in the baseline state, so an event usually appears in the $I(t)$ as an abrupt, short-duration spike or excursion from the baseline.

For example, DNA polymerase spends more time waiting for the ssDNA template strand and complementary nucleotide to arrive in the binding pocket than catalyzing the reaction (35, 36). When the correct substrates are aligned in the binding pocket, the polymerase changes conformation to the closed state to push the DNA strand to the next nucleotide position, and the corresponding signal measured by the SWCNT-FET switches to a second state (whose position is usually lower in current than the baseline state) (2). Control experiments confirmed that the switch to the second state only occurs when both the ssDNA template strand and complementary nucleotide are present in solution, which is correlated with the conformation change. Thus, the duration of the baseline state is the time the polymerase spends in the open conformation, waiting for the next nucleotide to arrive, while the duration of the second (lower) state is the time in the closed conformation.

The standard method for analyzing the durations of the states is by making a histogram of the calculated durations for each state to generate a probability distribution. Biomolecule activity typically follows a Poisson point process, in which events occur stochastically and independently but with some average rate λ and characteristic time $\tau = \frac{1}{\lambda}$ (37). In a Poisson process, the measured rate fluctuates in time with a standard deviation equal to the average: $r(t) = \lambda \pm \lambda(t)$. The probability distribution $P(t)$ of waiting times between events for a Poisson point process takes the form of an exponential distribution:

$$P(t) = \frac{1}{\tau} e^{-\frac{t}{\tau}} = \lambda\, e^{-\lambda t}$$

Thus, the characteristic time for a biomolecule's conformational change can be calculated by fitting an exponential function with no Y offset to the histogram and extracting τ from the fit. This is shown as a straight line on a semi-log plot. Often, the histogram of state durations deviates from a single-exponential shape, taking the form of a double-exponential or stretched exponential function. A double-exponential, appearing approximately piecewise-linear in a semi-log plot and exhibiting two characteristic times τ_1 and τ_2, suggests that the biomolecule alternates between two separate Poisson processes. A stretched exponential suggests that dynamic disorder causes the characteristic rate to change with time (38-40).

Additional analyses of the distributions of states is discussed in Chapter 5.

CHAPTER 2

Electronic Single-Molecule Measurements of the 3C6-Paclitaxel Binding Interaction

2.1 Introduction

Over the last 30 years, antibodies have become a ubiquitous tool for producing molecules, devices, and systems that possess specificity, natural or engineered, for a particular substance (41). Though medical researchers first pursued monoclonal antibody therapy for the treatment of various diseases (42), others soon began to use antibodies in applications ranging from simple selective linker molecules in ELISAs (43) to sensors that can detect specific molecular targets in a solution mixture (44) or even harmful pathogens (45, 46).

Antibodies, also called immunoglobins, are a type of protein produced by the immune systems of vertebrates (47) to assist in identifying and neutralizing pathogens, including viruses, bacteria, and a variety of toxins, by targeting characteristic molecules or structures on the pathogens called antigens. An individual antibody molecule, shown in the schematic in Figure 2.1, is shaped like a "Y", with two arms and a base (48). On each arm, there is one binding site at the tip (called the Fragment antigen-binding, or Fab, region), which is highly variable from one molecule to another, and which selective binds to a specific antigen. The base (called the Fc region) dictates the antibody's interaction with the rest of the immune system, including signaling for the appropriate response by the immune system, and is constant for each class of antibody. Clones of an antibody molecule with a specific Fab

variation can be used as a marker to tag a specific antigen. Though each antibody variant primarily targets one antigen in a "lock-and-key" interaction (49), recent studies showed that antibodies can interact and bind with antigens other than their original target (50, 51).

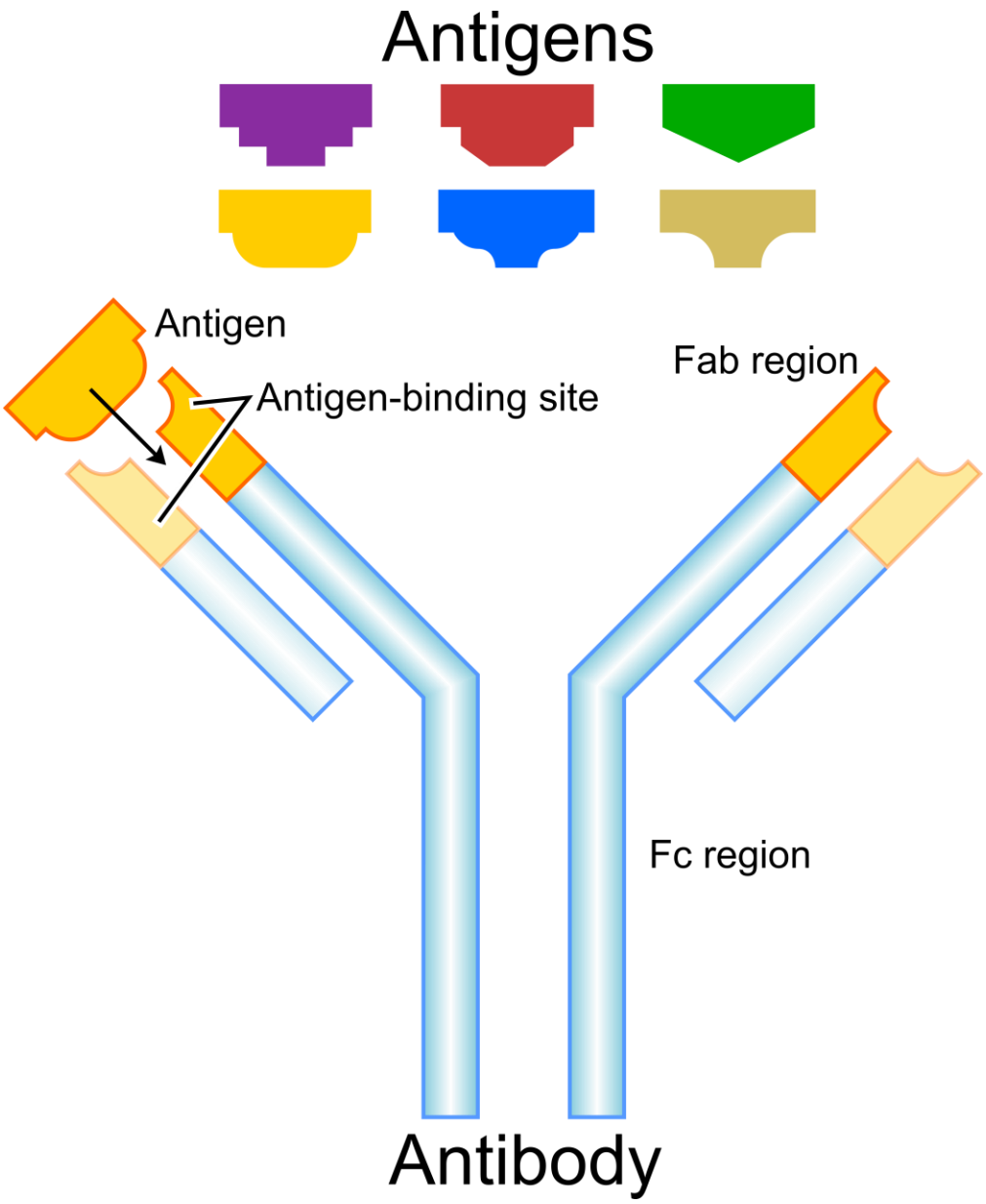

Figure 2.1: Schematic diagram of an antibody molecule, with the Fc region at the bottom in bluish-gray and the two Fab regions in yellow. Only one of the six antigens shown at the top can bind to the antibody's binding site.

Paclitaxel, also known by its trademark name Taxol®, is used for the treatment of breast, ovarian, lung, and other types of solid tumor cancers, and as such is included on the World Health Organization's List of Essential Medicines (52). Though an effective chemotherapy drug, paclitaxel's poor solubility in the bloodstream complicates drug delivery and dosage control, leading to variations in effective dosage and producing unnecessarily severe side effects (53). Many of these side effects are due to its delivery vehicle, Cremophor EL (54), a polyoxyethylated castor oil with its own toxic effects and which also forms micelles in the bloodstream, trapping paclitaxel inside (55, 56). In current treatment protocols, scheduled injections of paclitaxel are delayed if a patient demonstrates a certain number of adverse effects after a previous injection (57). Additional studies showed that keeping paclitaxel concentrations above 10 nM for over 24 hr activated mechanisms involved in cell death (58), suggesting that careful monitoring of paclitaxel concentrations near cancerous tissue might facilitate optimized dosing to reduce side effects while maintaining treatment effectiveness.

Experiments using the weakly-interacting monoclonal antibodies (3C6 and 8A10) show partial inhibition of paclitaxel's toxicity (59), demonstrating these antibodies' selectivity for the drug and their potential use as molecular sensors for paclitaxel concentration (60). However, commercially-available antibodies, even monoclonal antibodies, suffer from batch-to-batch variation (61, 62) in binding kinetics and affinity, which reduces the precision of antibody-based sensors. Such variation exists even between individual molecules, indicating a need for single-molecule studies of the antibody binding interaction to identify sources of variation.

Currently, studies of antibody-antigen binding often are performed as ensemble measurements, whether through surface plasmon resonance (SPR) (63-66), Raman spectroscopy (67), or nuclear magnetic resonance (NMR) spectroscopy (68). These techniques, though useful for characterizing antibody affinity and ensemble kinetics, do not have the sensitivity to observe the dynamics of an individual antibody molecule. Experiments with single molecule sensitivity utilize atomic force microscopy (AFM) (64, 69) or, more recently, single-molecule Förster resonance energy transfer (smFRET) (70, 71), which is dependent on fluorophores to generate its signal. Fluorophores commonly emit intermittently and with limited photon fluxes, restricting the time resolution that smFRET can achieve, and the fluorophores' tendency to photobleach places an upper limit on the observation time. With such limitations, smFRET cannot determine the presence or absence of sub-millisecond transient events or intermediate states.

In this chapter, the electronic measurement method described in Chapter 1 is used to record the binding activity of paclitaxel with antibody 3C6, which has some affinity for paclitaxel as an antigen. Ensemble and single-molecule measurements are utilized to obtain both time-averaged I-V_g curves and constant liquid-gate-potential, temporal current recordings ($I(t)$). Comparisons of the I-V_g curves do not reveal any consistent concentration-dependent shifts, but the $I(t)$ recordings show concentration-dependent rates of binding events. Analysis of single-molecule $I(t)$ recordings show that the 3C6-paclitaxel complex approximates a two-state system and exhibits concentration-dependent single-molecule binding kinetics that correlate to ensemble binding kinetics obtained in other experiments.

2.2 Experimental Methods

2.2.1 SWCNT-FET Device Preparation

SWCNT-FET devices were fabricated according to the procedure outlined in Section 1.2 and passivated with PMMA. The devices were electrically characterized in phosphate-buffered saline (pH 7.3, hereafter called PBS) solution (procedure outlined in Section 1.3) to obtain I-V_g curves and to test for RTS noise in the device. Any device demonstrating RTS noise before antibody attachment was discarded.

Ensemble SWCNT devices were created using CNTs from solution. Solubilized CNTs (NanoIntegris) were deposited onto a wafer by spin coating, leaving a dense film of randomly-oriented CNTs on the wafer surface. Deposition of metal electrodes on top of this film produced carbon nanotube network field effect transistors (CNTN-FETs). Subsequent electrical characterization showed that these devices were p-type transistors with lower on/off current ratios (~4-10) than the single-SWCNT-FETs, due to the presence of some metallic CNTs in the solution mixture. The CNTN-FETs were covered with PMMA, then portions of PMMA in the gap between electrodes were removed to expose the CNT network.

2.2.2 3C6 Antibody Attachment

40 μL of 1 mM pyrene-maleimide linker in ethanol was pipetted onto the top surface of a PMMA-covered chip, which was then left at room temperature (~22°C) for 30 min. The chip

was rinsed under a constant drip of 0.1% Tween-20 (MP Biomedicals) in ethanol for 5 s, then rinsed under flowing DI water for another 5 s, to remove any excess pyrene maleimide.

3C6 antibodies were attached to the SWCNTs in solution. Anti-SA2 antibody (Autotelic + Abcam (ab117725)), hereafter called 3C6, which is a monoclonal human IgG1 antibody, was diluted to 53 nM in PBS buffer, then divided into 60 μL aliquots and frozen for storage. When needed, the 3C6 aliquot was thawed and pipetted onto the top surface of the chip, and the chip was left at room temperature (~22°C) for 30 min. This concentration was chosen to facilitate, on average, one attachment per SWCNT. The chip was rinsed under a flowing solution of 0.1% Tween-20 in PBS solution for 5 s, then rinsed under flowing DI water for another 5 s, to remove any excess 3C6 from the chip surface. The chip was submerged in PBS solution at room temperature (~22°C) for short-term storage.

2.2.3 Electrical measurements

The target antigen, paclitaxel (Autotelic), was dissolved in solution by pipetting small amounts of the storage solution (6 mg/mL of paclitaxel dissolved in polyoxyethylated castor oil and dehydrated alcohol) into PBS buffer to produce the target concentrations. Paclitaxel is considered toxic, so any procedures involving the concentrated storage solution were performed in a fume hood with the appropriate safety measures. Fresh paclitaxel solutions (prepared within 2-3 hours) were used in each measurement, since paclitaxel degrades in aqueous solutions (72).

Measurements were then performed in approximately logarithmically-spaced concentrations of paclitaxel in PBS solution ranging from 20 pM to 200 nM. For each concentration, a I-V_g curve (0.6 V < V_g < 0.4 V) was first acquired, obtaining time-averaged measurements with each point being an average of 0.5 s of data at 5 kHz. Such I-V_g measurements correspond to those performed by other CNT biosensor experiments (27, 73). Then, the V_g was held constant, and $I(t)$ was measured for a minimum of 300 s at an acquisition rate of 100 kHz. After completing recordings for a particular solution, the chip would be rinsed under flowing DI water for 5 s, then submerged in PBS solution until the next measurement. After all electrical measurements were completed, the device was rinsed under running DI water for 5 s and then dried with a compressed air gun.

2.2.4 Atomic Force Microscopy Imaging

Atomic force microscopy (AFM) (Pacific Nanotechnology Nano-R) was used to determine how many 3C6 antibodies were attached to the SWCNT. Attached 3C6 appeared as dots, ~3 nm high, overlapping with the line of the SWCNT, which is consistent with the approximate size of the 3C6 molecule (64). The AFM images were used to categorize devices according to the number of SWCNTs and the number of attached 3C6 molecules. The categories were: single-SWCNT and single-3C6 (single-molecule device, Figure 2.2A), single-SWCNT and multiple 3C6 (few-molecule device, Figure 2.2B), and multiple-SWCNT and multiple 3C6 (ensemble device, Figure 2.2C).

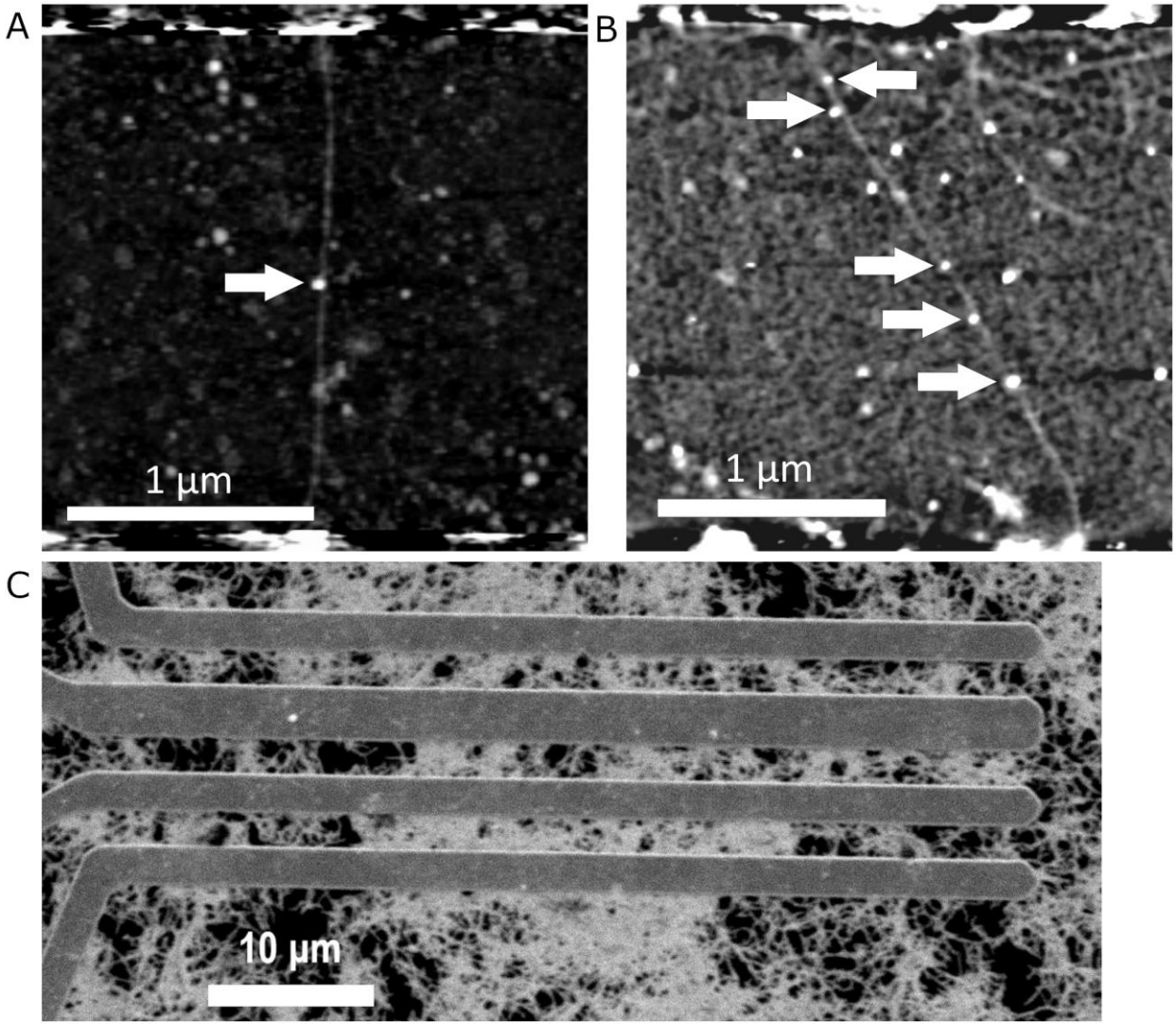

Figure 2.2: (A) Atomic force microscopy (AFM) image of a single-molecule device. The individual 3C6 molecule appears as a dot (shown with an arrow) overlapping with the line of the SWCNT, here shown running down the middle of the image. (B) AFM image of a few-molecule device, with arrows pointing to multiple attached 3C6 molecules. (C) Scanning electron microscopy (SEM) image of several ensemble devices, where each device is composed of two adjacent electrodes (gray) connected by a network of CNTs (white). The individual 3C6 molecules are not visible with this imaging technique.

2.2.5 Signal Analysis

The obtained I-V_g curves were normalized to allow comparison of curves resulting from different paclitaxel concentrations and to quantify the resulting shifts. First, the I-V_g curves

were converted to G-V_g curves by dividing I by V_{sd}. Then, a constant was subtracted from the G-V_g curves to align all the curves at the right-most end (at V_g = +0.4 V). This was especially important for comparing the G-V_g curves of the ensemble devices, since the CNT networks contained some metallic paths that continued to conduct even when all of the semiconducting SWCNTs were gated off (74). G-V_g curves from the same device were plotted together, and changes in maximum current (measured at V_g = -0.6 V) and shifts of the threshold voltage were calculated.

The raw $I(t)$ signal was converted to a simple binary signal to facilitate two-state analysis. The $I(t)$ signal was converted to conductance using the relation $G(t) = I(t)/V_{sd}$. The total > 300s data set was broken up into short segments of ~1-3s. Segments were detrended by subtracting the 1/f noise ($\Delta G(t) = G(t) - \langle G(t) \rangle$, where the average is calculated over 100 μs), and then filtered using an implementation of the NoRSE algorithm (75). Each data point in the filtered signal was assigned to either the high or low state based on a simple threshold algorithm, with the threshold approximately placed halfway between the baseline and the highest-magnitude RTS signal. The threshold algorithm also ignored any events that were shorter than 50 μs to reject spurious event-like noise peaks that were too short to have been amplified by the preamplifier. An example of the binary signal superimposed on the detrended $\Delta G(t)$ signal, showing the accuracy of fitting, are displayed in Figure 2.3.

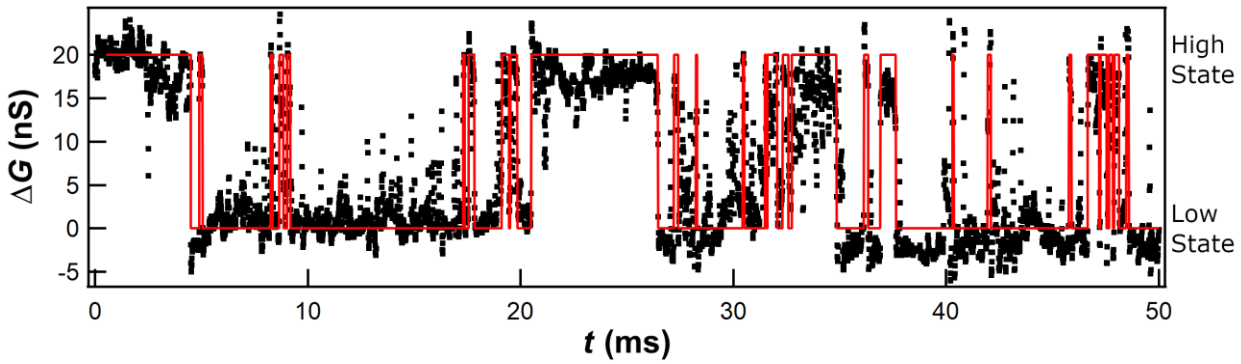

Figure 2.3: Plot of 50 ms of the detrended conductance $\Delta G(t)$ (black) for a typical single-molecule measurement (in this case at 70 nM paclitaxel), with the calculated binary signal shown in red. The conductance jumps between two states (indicated by the two levels of the binary signal). Time durations of individual states are represented by the length of each horizontal line segment.

The duration for each occurrence of the high and low states was calculated. For each concentration and each state, the distribution of the state durations was plotted on a semi-logarithmic histogram, then fitted to a bi-exponential function with zero Y offset. Figure 2.4 shows an example of the probability distribution, and the corresponding bi-exponential fit, for both the high state and low state. The fitted function appeared approximately piecewise-linear in a semi-log plot, corresponding to a bi-exponential distribution with two characteristic times τ_1 and τ_2. In addition, the arithmetic average time $\langle t \rangle$ of each state was calculated. The average probability of time that the $\Delta G(t)$ signal spent in the high state versus the low state was calculated by normalizing the binary signal to have zero offset and unity amplitude, then taking the average of the normalized signal to obtain the high state average. The number of switches (defined as the number of changes from the low state to the high state) was calculated from the binary signal, and the average switching rate calculated for each concentration.

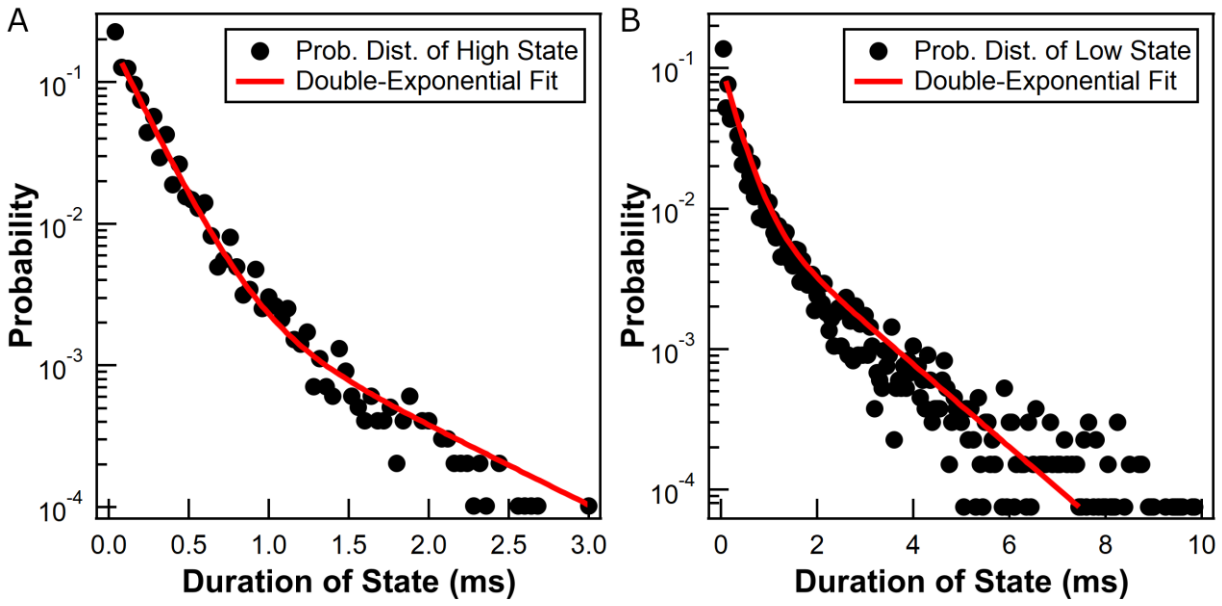

Figure 2.4: Typical log-linear plots of the probability distributions (black) for both the high state (A) and the low state (B), with the double-exponential fit for each state shown in red. Double-exponential functions appear as approximately piecewise-linear functions with two linear regions corresponding to the two characteristic values, in this case the time constants τ_1 and τ_2.

2.3 Results

Every device showed an immediate and irreversible shift in the G-V_g curve upon the first exposure to paclitaxel. The single-molecule devices gradually equilibrated over 60 seconds, and ensemble devices gradually equilibrated over 120 seconds. The shifts were most visible and uniform for the single-molecule device, for which the electrostatics of the attached 3C6 were most predictable. For these devices, the curve uniformly dropped to ~80% of the pre-exposure value over the entire V_g range (Figure 2.5A). The curves for the few-molecule devices, on average, increased to ~110% of the pre-exposure values while also shifting by approximately -100 mV, causing the current to decrease for most V_g < -0.4 V and increase for -0.4 V < V_g < -0.1 V (Figure 2.5B). The response of ensemble devices was least predictable

(Figure 2.5C). Attempts to rinse away the paclitaxel, however, had negligible effect on the G-V_g curves, and subsequent exposures never produced a response as large as the initial exposure. Furthermore, these subsequent shifts were uncorrelated with paclitaxel concentration. Figure 2.5C shows an example of a non-monotonic response, where initial exposure of paclitaxel increased G, but subsequent increases in paclitaxel concentration shifted G in an uncorrelated manner.

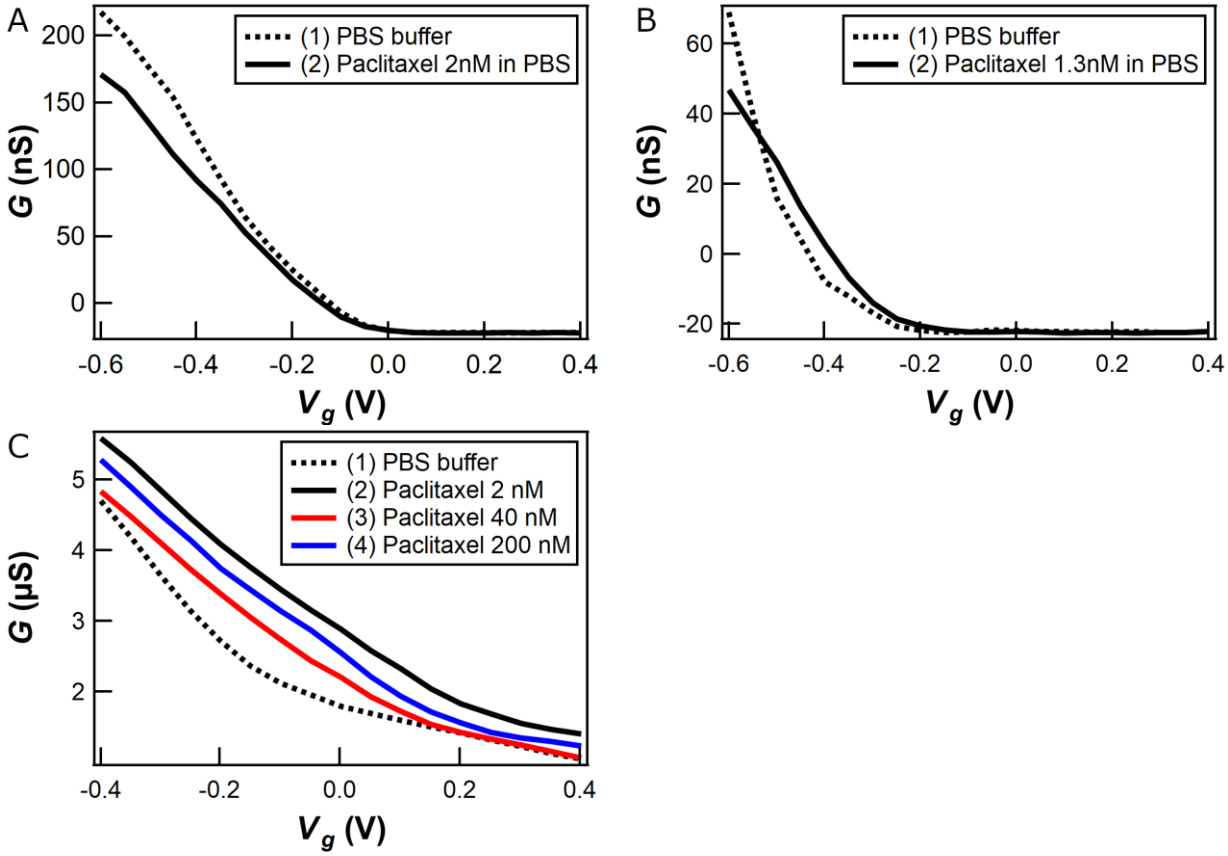

Figure 2.5: Typical G-V_g curves for a (A) single-molecule device, (B) few-molecule device, and (C) ensemble device, with measurements in buffer shown by a dotted black line and measurements with paclitaxel shown as various colors of solid black lines. The number in parenthesis in each plot legend indicates the order of the measurement.

The shift in the G-V_g curve was also accompanied by single-molecule fluctuations $\Delta G(t)$ associated with paclitaxel binding and unbinding. Figure 2.6 shows typical graphs of $\Delta G(t)$ for the three types of devices. The single-molecule devices gave the simplest signals consisting of stochastic fluctuations between only two states. This two-level switching was consistent with the presence of a single 3C6 binding site and also with previous experiments labeling SWCNTs with single biomolecules (1, 2). The few-molecule devices displayed combinations of two-level, three-level, or multi-level fluctuations that were consistent with one or more active 3C6 antibodies. Assuming each 3C6 molecule is independent, multiple 3C6 molecules should contribute additively to the $G(t)$ signal, so these multi-level signals display the activity of multiple 3C6 molecules. In the extreme case of ensemble devices, $\Delta G(t)$ signals no longer displayed clearly-resolvable states, most likely because of the averaging effects of multiple active sites. With all the devices, $\Delta G(t)$ fluctuations disappeared when the paclitaxel was rinsed away, even though the shifts in G-V_g curves were not entirely recovered. The difference suggests that part of the shift was caused by nonspecific binding to the surface, whereas the $\Delta G(t)$ fluctuations were driven by specific binding to the 3C6 antibodies.

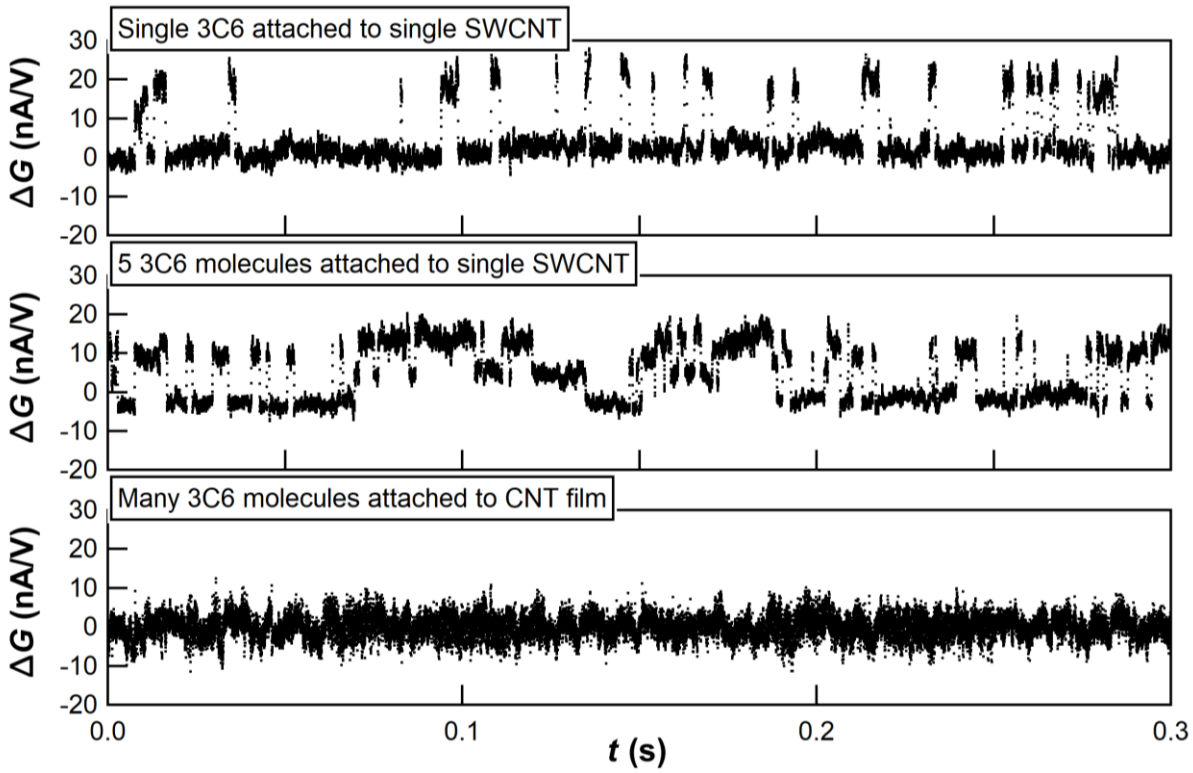

Figure 2.6: Plots of typical $\Delta G(t)$ recordings for single-molecule (top), few-molecule (middle), and ensemble (bottom) devices. Single-molecule devices showed clear two-level switching, few molecule devices sometimes showed multi-level switching, and ensemble devices showed no clear switching.

Each discrete level in $\Delta G(t)$ was interpreted as being caused by one active 3C6. However, the number of levels did not exactly match the number of 3C6 attachments observed by AFM. Sometimes, devices with 2-4 3C6 molecules produced a simple, RTS signal consistent with only 1 active binding site. Other few-molecule devices produced complex $\Delta G(t)$ signals with multiple levels or complex noise. This demonstrated that not all 3C6 molecules were active and able to produce an electrical response to paclitaxel binding. Further analysis of the $\Delta G(t)$ data were performed only with the single-molecule devices which produced only 2-level RTS, which are the most easily-interpreted signals.

Unlike in previous research with single-cysteine mutants, the 3C6 molecule (a type IgG1 immunoglobin) has multiple possible attachment positions due to the many cysteine residues exposed on the surface (76, 77). Figure 2.7 shows the ribbon diagram of an IgG1 molecule (78), with the individual cysteine residues highlighted in red. Though all of the cysteine residues in an IgG1 molecule are nominally engaged in disulfide bonds, other experiments showed that many of the cysteines could be uncoupled from its disulfide bond pair (77, 79), especially under partially denaturing conditions, and therefore could be available for maleimide-thiol conjugation. All measurements in this work were performed with wild-type (unmutated) 3C6 molecules, with which explicit control of the attachment site and orientation is impossible. Thus, these measurements provided very little information regarding the origin of the conductance fluctuations, including the position of the most influential residues or the conformational motions responsible for the signal. In addition, some of these attachment sites might have anchored the 3C6 molecule in an orientation which prevented electrical gating of the SWCNT.

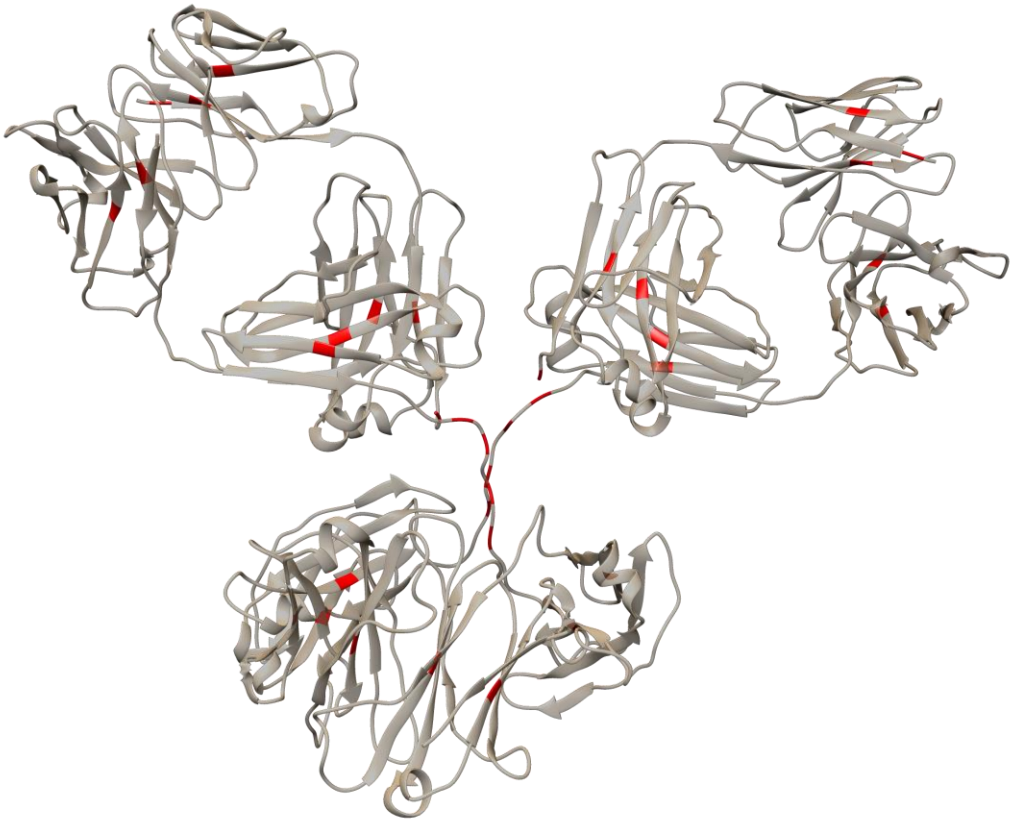

Figure 2.7: Ribbon structure of an IgG1 molecule, such as 3C6, oriented like a letter "Y", with two binding sites at the top left and right. The cysteine residues are highlighted in red, showing the possible locations for the maleimide-thiol linkage to the SWCNT.

The simple two-level $\Delta G(t)$ signals from the single-molecule devices were analyzed in terms of bound and unbound states. Figure 2.8 displays, on the left, $\Delta G(t)$ signals from 6 different measurements, corresponding to no paclitaxel (buffer only, top) and 5 concentrations of paclitaxel. The right side of the figure zooms in on the data highlighted in the red box and shows the calculated bound and unbound states as the two horizontal levels of the red lines. Qualitatively, at lower paclitaxel concentrations, the conductance was mostly in the lower state, with occasional, short jumps into the higher state. At higher concentrations, the behavior was exactly opposite, residing mostly in the higher state and occasionally jumping into the lower state. At concentrations between 10 and 70 nM, the conductance exhibited

stochastic, high frequency switching between states and spent substantial time in both states. The lower state was interpreted as the unbound state, using measurements in low concentration, in which the 3C6 molecule spent most of its time waiting for a paclitaxel molecule to arrive. The high state was interpreted to be the bound state, using measurements in high concentration when the 3C6 molecule was surrounded by an abundant source of paclitaxel.

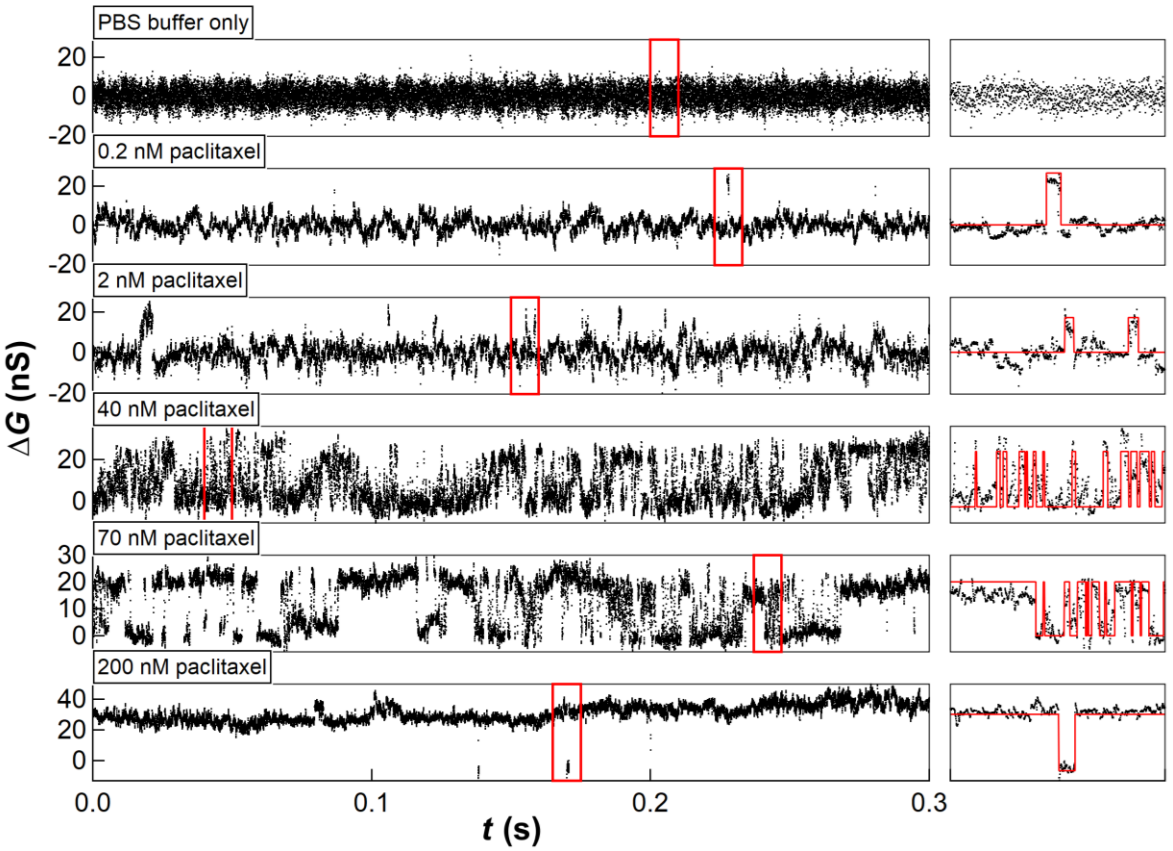

Figure 2.8: Plots of typical $\Delta G(t)$ recordings for a single-molecule device at various paclitaxel concentrations. The plots on the left are 0.3 s of data, and the plots on the right are expanded views of the corresponding regions highlighted by the red boxes (each 0.01 s long). The calculated binary signal for each expanded view is shown in red.

Transitions between events were stochastic, approximating a Poisson point process with a time-dependent rate constant $\tau_P(t)$. At a given moment, the distributions of time durations for both the bound and unbound states were approximately Poisson (approximately fit by an exponential function with a characteristic time τ), but the instantaneous binding rate varied stochastically. Nevertheless, long data records could be accumulated, and the resulting accumulated distribution of the time durations for both the bound and unbound states formed either single or bi-exponentials. Each distribution was described according to two parameters. One was the Poisson time constant τ, which could be considered the most probable binding rate. Since many of the distributions formed bi-exponentials, there were often two τ values for each state. The other was the arithmetic mean of time durations $\langle t \rangle$, which incorporated all of the actual deviations from Poisson behavior and could be correlated with the average binding rate as observed by ensemble measurements.

Figures 2.9A and B show the dependence of the four τ values on paclitaxel concentration. Quantitatively, both $\tau_{unbound1}$ and $\tau_{unbound2}$ decreased consistently over the entire concentration range in a manner approximating a power law, which form straight lines in a log-log plot. $\tau_{unbound1}$ dropped from $\sim 10^{-1}$ to $\sim 10^{-4}$ s, with an exponent of $\sim$-0.56. $\tau_{unbound2}$ dropped from $\sim 10^{-2}$ to $\sim 10^{-5}$ s, with an exponent of $\sim$-0.71. The fractional exponents obtained here differed from the exponent of unity expected for a second-order reaction in basic receptor–ligand kinetics. This suggested the presence of intermediate steps in the paclitaxel-3C6 binding process, not yet observed, that alter the reaction kinetics to produce the observed fractional reaction order.

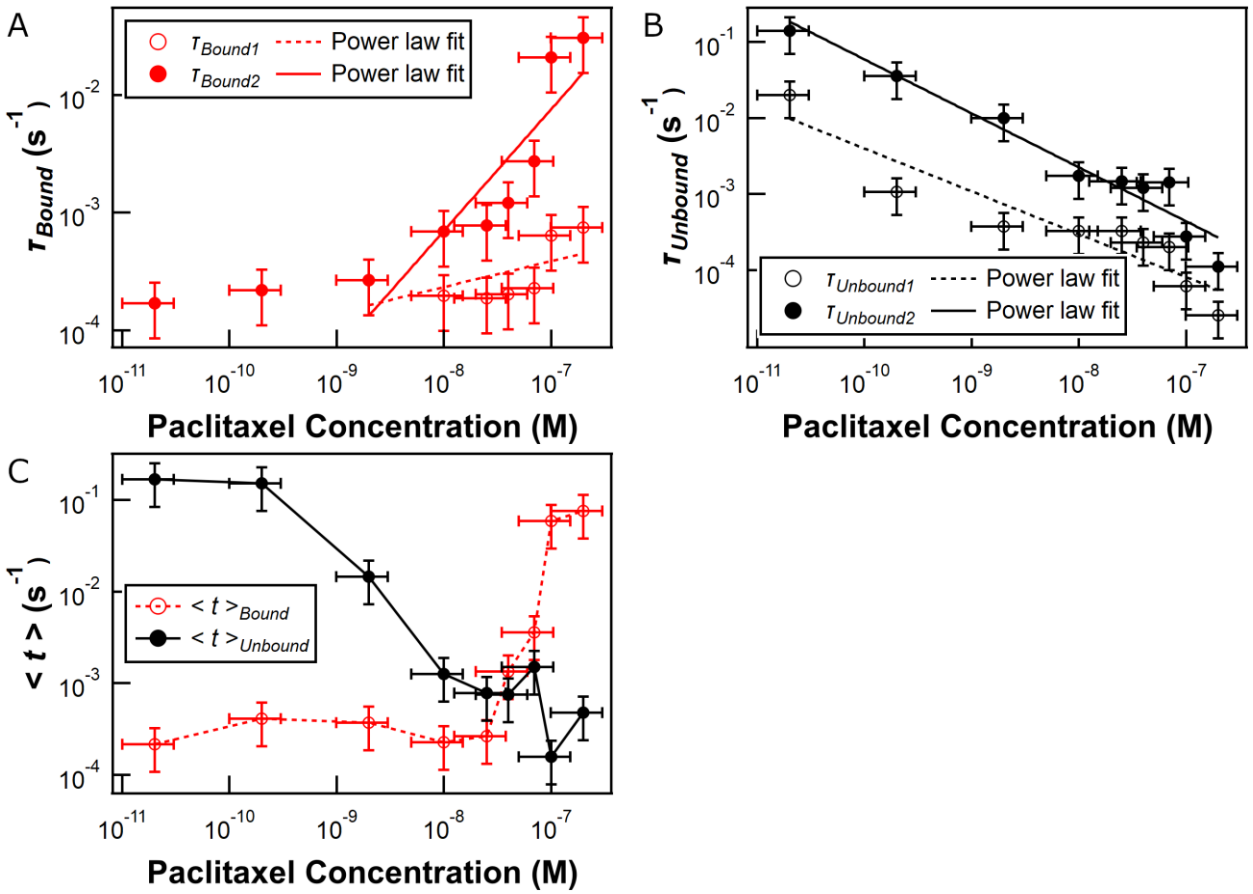

Figure 2.9: (A) Plot of the dependence of the two $\tau_{unbound1}$ values on paclitaxel concentration. The first value is chosen to always be the larger of the two values. The power-law fits for each value is shown, with the fit for $\tau_{unbound1}$ having a slope of ~-0.56 and the fit for $\tau_{unbound2}$ having a slope of ~-0.71. (B) Plot of the dependence of the two τ_{bound1} values on paclitaxel concentration. The power-law fits for each value is shown, with the fit for τ_{bound1} having a slope of ~+1.0 and the fit for $\tau_{unbound2}$ having a slope of ~0.22. Below 2 nM, the two τ_{bound1} values are identical and approximately constant. (C) Plot of the dependence of <*t*> on paclitaxel concentration.

By contrast, τ_{bound1} and τ_{bound2} demonstrated two distinct regions of concentration dependence. At paclitaxel concentrations below 2 nM, both τ_{bound1} and τ_{bound2} remained approximately constant at ~10^{-4} s. At these concentrations, the resulting distribution of τ_{bound} showed no evidence for a bi-exponential, so the single exponential τ_{bound} was used for both. As paclitaxel concentration increased, the τ_{bound} values separated, with τ_{bound1} remaining relatively constant and τ_{bound2} rising to ~10^{-2} s. In this separated region, τ_{bound1}

demonstrated a power law concentration dependence with exponent ~1.0, while τ_{bound2} demonstrated an exponent of ~0.22. From the definition of the ensemble dissociation constant, $K_d = \frac{1}{K_a} = \frac{k_{off}}{k_{on}}$, and by assuming that k_{off} was related to the constant τ_{bound} value through $k_{off} = \frac{1}{\tau_{bound1}} = 10^4\ \mathrm{s}^{-1}$ (80, 81), the on rate constant was estimated to be $k_{on} = \frac{k_{off}}{K_d} = \frac{10^4\ \mathrm{s}^{-1}}{10\times10^{-9}\ \mathrm{M}} = 10^{12}\ \mathrm{s}^{-1}\mathrm{M}^{-1}$.

Figure 2.9C shows that $\langle t \rangle_{bound}$ and $\langle t \rangle_{unbound}$ exhibited opposing concentration dependence, each mirroring the other around some crossover concentration value. At paclitaxel concentrations below 10 nM, $\langle t \rangle_{bound}$ stayed on the order of 10^{-4} s while $\langle t \rangle_{unbound}$ was roughly 10^{-1} s. At high concentrations, the $\langle t \rangle_{bound}$ and $\langle t \rangle_{unbound}$ values switched in orders of magnitude, such that $\langle t \rangle_{bound}$ rose to around 10^{-1} s and $\langle t \rangle_{unbound}$ dropped to the order of 10^{-3} s. The greater value of $\langle t \rangle$ as compared to τ reflected the more frequent occurrence of longer events than predicted by a purely Poisson point process. These deviations reflected the presence of brief pauses in binding activity, perhaps due to the 3C6 molecule becoming stuck in some conformation that prevented binding or unbinding. The presence of such pauses could be due to dynamic disorder in the 3C6 molecule conformation that form a collection of substates, perhaps driven by thermal fluctuations (38-40), or else could be caused by interactions between the 3C6 molecule and the surface of the SWCNT device. At paclitaxel concentrations around 30 nM, both $\langle t \rangle_{bound}$ and $\langle t \rangle_{unbound}$ converged at about 10^{-3} s, indicating some sort of balance between binding and unbinding. This crossover concentration value was in approximate agreement with the value of $K_d = 10$ nM (60), as found through ensemble measurements of the 3C6-paclitaxel binding interaction.

Figure 2.10 shows that the time-averaged binding probability was concentration dependent and indicated a similar crossover concentration value. At low concentrations (< 200 pM), the signal resided in the low state for over 99% of the time, whereas at high concentrations (> 100 nM), the signal resided in the high state for over 99% of the time. Between these two extremes, the probability of the high state increased with concentration in a manner approximating a logarithmic sigmoidal binding curve, also known as the Hill-Langmuir equation. A fit to such a curve, given by:

$$\text{Probability} = \frac{1}{1 + \left(\frac{K_d}{[L]}\right)^n}$$

is shown as solid lines. The resulting fit showed a 50% high state / 50% low state crossing at ~30 nM, fairly close to the ensemble-measured $K_d = 10$ nM, and also a Hill coefficient of ~1.8. The correlation between the time-averaged binding probability calculated here and the ensemble-measured K_d value for the antibody 3C6-paclitaxel interaction suggested that the ensemble average reflected a simple time-average of the single-molecule binding dynamics for this particular interaction, and that the binding process itself was ergodic.

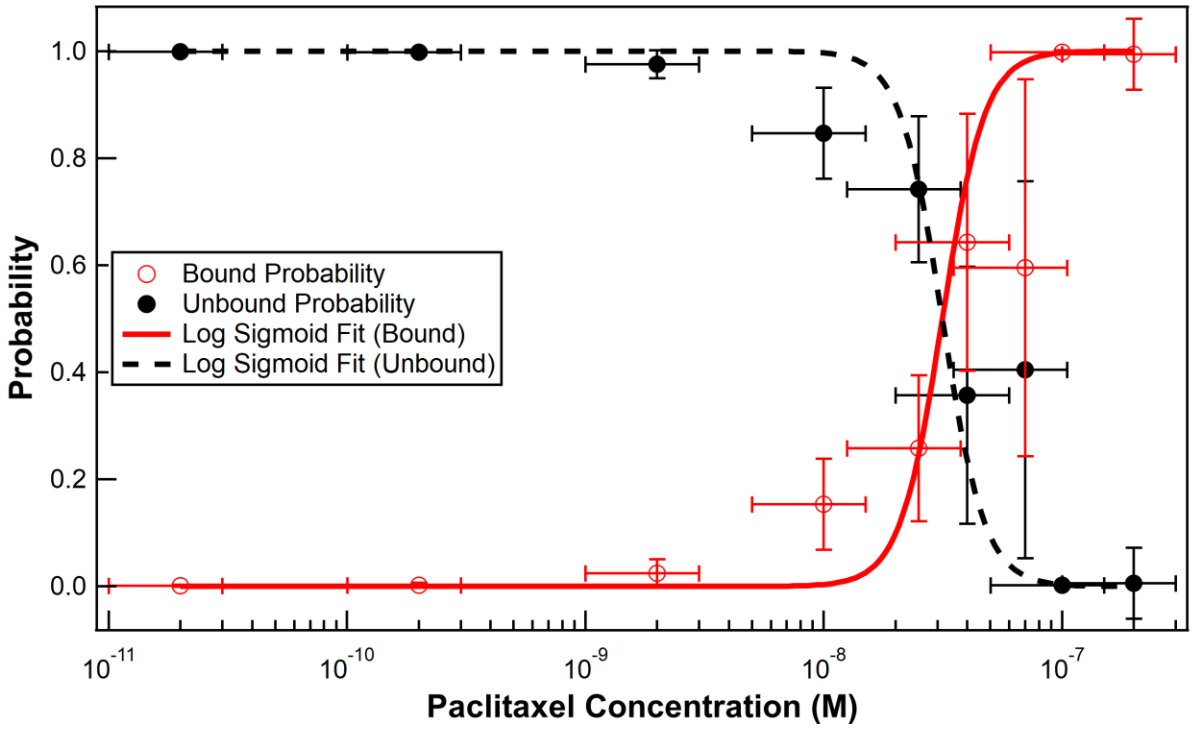

Figure 2.10: Plot of the overall probability of the bound (high) state (black) and the unbound (low) state (red) as a function of paclitaxel concentration. The corresponding fits to a logarithmic sigmoidal binding curve are shown as solid lines. The 50% bound / 50% unbound crossover occurs at ~30 nM. The Hill coefficient from the fit is ~1.8.

Deviations from the exponential fit occurred only at longer time durations, which occurred much less frequently than the short events, and such deviations described events that were longer than what was predicted by the fit. The presence of some very long event times ($\tau >$ 1 s) supported the hypothesis that the 3C6 became stuck in one conformation or otherwise became inactive for certain periods of time.

2.4 Discussion

In these experiments, changes in the G-V_g characteristics were uncorrelated with paclitaxel concentration for all device types (single-molecule, few-molecule, or ensemble). At first glance, the lack of correlation in the response of the ensemble devices would seem to contradict the results of similar experiments performed on graphene transistors (28, 32). In both graphene transistors and ensemble CNT devices, there are many conducting paths for electron flow which are simultaneously modulated by many analyte-transistor interactions, such that changes in G-V_g itself serves as the signal for detecting analyte concentration. In addition, other biosensing experiments using single-CNT transistors (27, 73, 82) showed direct correlations between the analyte concentration and the absolute value of the equilibrium source-drain conductance, which was not observed in this work. Instead, only the $\Delta G(t)$, fluctuations away from the baseline conductance, was directly related to paclitaxel concentration.

One possible reason for this discrepancy was contamination or nonspecific adsorption of paclitaxel on surfaces surrounding the 3C6-SWCNT device. Both paclitaxel and its storage solution, castor oil, are generally insoluble in water (55) and were difficult to reliably rinse away. Once either substance came into contact with the 3C6-SWCNT device, it likely coated the surface of the CNT sidewall, irreversibly altering the equilibrium electrical characteristics of the device. This shows that the antibody-mediated binding interaction on the SWCNT device is robust against some surface contamination even when the remainder of the CNT device becomes contaminated. Also, the equilibration time of the single-molecule

devices (under 60 s) was much less than the equilibration time for the other CNT- or graphene-based sensors in the studies mentioned above. Thus, these SWCNT-antibody devices could be used as robust sensors of binding interactions or for cross-checking measurements using other techniques when contamination or low-purity samples might otherwise corrupt a measurement.

The bi-exponential fits of the unbound time distributions showed that the mechanism of paclitaxel binding to the 3C6 molecule involves two independent Poisson-like processes. The similar concentration dependence of both $\tau_{unbound1}$ and $\tau_{unbound2}$ suggests that both processes correspond to actual 3C6-paclitaxel binding events. Perhaps the second process corresponds to paclitaxel attaching to the second antigen binding site on the 3C6 molecule, since all antibody molecules possess two binding sites in separated domains. Despite the two binding sites, the 3C6 molecule did not produce three current levels, which is unlike previously-studied enzymes with multiple ligand binding sites, such as protein kinase A (3). This may be because the SWCNT-FET biosensor is sensitive to conformational changes within ~1 nm of the CNT sidewall (25). Since the attachment location for the specific 3C6 molecule studied here was unknown, it is possible that the SWCNT biosensor in this case was linked to one of the arms of the 3C6 molecule. Due to the inherent flexibility of antibodies, this would leave the SWCNT directly sensitive to the electrostatics of one binding site and not the other, but the effect of the distant binding site might still be detected through the altered kinetics of the local binding site.

Antibody-antigen interactions are expected to follow general receptor-ligand kinetics, which treat binding as a second-order reaction and unbinding as a first-order reaction. In this work, the approximately power-law concentration dependence of both $\tau_{unbound}$ values were somewhat similar to a second-order reaction, but there were significant deviations. In particular, a second-order reaction rate should depend on ligand concentration as $[L]^1$, which means that $\tau_{unbound}$ should depend on ligand concentration as $\tau_{unbound} = \frac{1}{k_{on}} \propto [L]^{-1}$. However, the power law dependence for $\tau_{unbound}$ values were much lower in magnitude than unity. The greatest deviations from the power law occured in the range of 2 nM and 70 nM, which was the same range over which the $\Delta G(t)$ signal displayed the most transitions from the unbound to bound state and back. The high rate of transitions resulted in a signal that was difficult to cleanly separate into two states, especially with the limitation of the ~12.5 kHz bandwidth introduced by the current preamplifier. In fact, the $\tau_{unbound}$ values corresponding to the high rates of transitions seemed almost to become constant at approximately 10^{-4} s, suggesting that any $\tau_{unbound}$ values smaller than this were not accurately resolved by the measurement setup used in this work. This, in turn, suggests that the true values for $\tau_{unbound}$ at concentrations above 40 nM should be lower than what was measured. Measuring $\Delta G(t)$ from 3C6-paclitaxel binding with larger bandwidth would facilitate more accurate measurements of lower $\tau_{unbound}$ values and likely reveal the true $[L]^{-1}$ dependence.

The approximately constant value of τ_{bound} for concentrations below K_d corresponded to a concentration-independent unbinding rate, matching the kinetics of a first-order reaction. At these concentrations, individual binding event were well-separated and easily distinguishable from each other, facilitating accurate measurements of the binding kinetics.

However, as paclitaxel concentration increased above K_d, the binding kinetics of the 3C6 molecule were too fast to accurately resolve, due to the limited measurement bandwidth. Thus, several individual binding and unbinding events were counted as one event, resulting in a τ_{bound} that measured the time of multiple binding events instead of one individual event, which explains the increase in τ_{bound2} with increasing paclitaxel concentration beyond K_d. In fact, at the highest paclitaxel concentrations measured (100 nM and 200 nM), the signal resided in the bound state most of the time, with very few transitions between bound and unbound states appearing in the signal, even though the theory of simple receptor-ligand kinetics would predict many fast binding and unbinding transitions at those concentrations. This shows that measurements at high ligand concentrations are limited by measurement bandwidth, and that repeating these measurements with larger bandwidth would produce a constant τ_{bound} for a wider range of concentrations.

The calculated binding and unbinding rate constants for the 3C6-paclitaxel interaction were eight orders of magnitude faster than the rate constants observed for other antibodies in experiments using ligand binding assays. The 3C6-paclitaxel interaction was specifically chosen for this experiment because its binding kinetics were predicted to be in a range which the SWCNT-FET device could measure. The SWCNT-FET device had an approximate bandwidth of $10^1 - 10^5\ \mathrm{s}^{-1}$, corresponding to $k_{off} \sim 10^1 - 10^5\ \mathrm{s}^{-1}$, which matched the kinetics of the 3C6-paclitaxel system ($k_{on} = 10^{12}\ \mathrm{M}^{-1}\mathrm{s}^{-1}$, $k_{off} = 10^4\ \mathrm{s}^{-1}$). By contrast, many antibodies exhibiting $K_d = 10$ nM that were measured by typical ligand binding assays demonstrated $k_{on} \sim 10^4 - 10^5\ \mathrm{M}^{-1}\mathrm{s}^{-1}$ and $k_{off} \sim 10^{-4} - 10^{-3}\ \mathrm{s}^{-1}$ (83, 84), kinetics slow enough for SPR or fluorescence but outside the range of SWCNT-FET devices. Thus, ligand

binding assays are complementary methods to the SWCNT-FET method used here, since each method measures different timescales. $k_{on} \sim 10^4 - 10^5\ \mathrm{M^{-1}s^{-1}}$

The effective K_d obtained in this experiment was 30 nM, slightly different from the ensemble-measured $K_d = 10\ \mathrm{nM}$. This discrepancy was expected, given that previous studies showed some variation in equilibrium and kinetic binding characteristics between different molecules of the same protein (61). In addition, paclitaxel was shown to degrade by hydrolysis in mildly basic solutions (72), which would lower the actual concentration of paclitaxel in solution during an experiment. Therefore, what was nominally 30 nM paclitaxel could actually have been closer to 20 nM, reducing the discrepancy between the single-molecule value reported here and the ensemble-measured value.

The value of 1.8 obtained for the Hill coefficient, close to the theoretical limit of 2 which corresponds to the number of binding sites, indicates a strong interaction between the two paclitaxel-binding sites on the 3C6 molecule. The strong cooperativity enhances paclitaxel binding to the second site even at low concentrations by making the second binding more favorable, which has been suggested by several studies (85-87). Cooperative binding could be the reason for the bi-exponential distribution of unbound times, resulting from paclitaxel binding both to the first and then to the second arms of the 3C6 molecule. As mentioned above, the SWCNT biosensor is only sensitive to conformational changes within ~1 nm of the CNT sidewall, limiting the sensitivity of the device to one binding site. Thus, the SWCNT biosensor could be indirectly sensitive to intra-antibody cooperativity.

2.5 Summary

The results presented here demonstrate the ability of SWCNT-FET biosensors to probe microsecond-scale details in binding kinetics, distinguish between mechanisms in receptor-ligand binding, and to measure ligand concentration in solution. Using this CNT biosensor, the 3C6-paclitaxel system was shown to exhibit binding kinetics approximating typical receptor-ligand binding, but accurate determination of the kinetic parameters was reduced by the limited bandwidth of the measurement. The results of this experiment suggest that greater accuracy could be obtained in experiments performed at higher bandwidth. Further analysis of 3C6-paclitaxel binding dynamics strongly suggested that the two arms of the 3C6 antibody molecule exhibited cooperative binding behavior. This work provides a new way to investigate rapid receptor-ligand interactions and may lead to a deeper understanding of the mechanics behind receptor-ligand specificity and the interactions between multiple binding sites on the same molecule.

CHAPTER 3

Electronic Single-Molecule Measurements of φ29 DNA Polymerase

3.1 Introduction

DNA polymerases are enzymes that synthesize double-stranded DNA (dsDNA) from single-stranded DNA (ssDNA) by incorporating complementary nucleotides into the nascent second strand, one nucleotide at a time. DNA polymerases are the linchpin in DNA replication, serving as the essential enzyme and proof-reader that both synthesizes the second DNA strand and checks for misincorporations. Due to their ability to discriminate between bases and selectively incorporate complementary nucleotides, DNA polymerases have been extensively studied, both to understand the mechanisms of catalysis (35, 88, 89) and proof-reading (89, 90) as well as for applications in DNA sequencing (91-94). Two specific DNA polymerases (Taq DNA polymerase and φ29 DNA polymerase) serve as the workhorses for the DNA sequencing industry (93, 94). Investigations of these two polymerases have utilized both ensemble techniques such as mutagenesis (95-101), crystallography (102-104), fluorescence (105, 106), and DNA amplification (107-110) as well as single-molecule techniques such as optical tweezers (111-114) and nanopores (115-120).

Like most DNA polymerases, φ29 DNA polymerase is nominally shaped like a right hand (with thumb, fingers, and palm subdomains, among others, as shown in Figure 3.1) (103, 104), with the catalytically active site on the inside of the palm domain. The enzyme exhibits

two main conformations: an open conformation which allows both ssDNA template and nucleotides to bind to the thumb and fingers domains, respectively, and a closed conformation which is correlated with nucleotide incorporation and translocation of the ssDNA strand. The enzyme's processing rate is temperature-dependent, ranging from ~5 nucleotides/s at 4°C to ~40 nucleotides/s at 30°C (109, 121). The exonuclease domain (gold color in Figure 3.1) is responsible for cleaving off any mismatched nucleotides (122), which contributes to the enzyme's fidelity (107). The structure of φ29 DNA polymerase is unique in possessing a TPR2 domain (magenta color in Figure 3.1), which forms a loop with the palm and thumb domains to keep the enzyme wrapped around the DNA strand. This domain allows the enzyme to serve as its own helicase to unwind dsDNA into individual ssDNA strands, unlike many other replicative polymerases which require separate helicase proteins. Due in part to this domain, φ29 DNA polymerase is one of the most processive polymerases, able to catalyze >70,000 consecutive base pairs before disassociating from the DNA strand (123). The enzyme is also a reliable and accurate replicative polymerase, with an error rate of 10^{-5} - 10^{-6} errors/dNTP (108). These last two characteristics of φ29 DNA polymerase make it an ideal enzyme for DNA sequencing, and several DNA sequencing platforms take advantage of these properties for high-throughput or long-read sequencing (91, 124, 125).

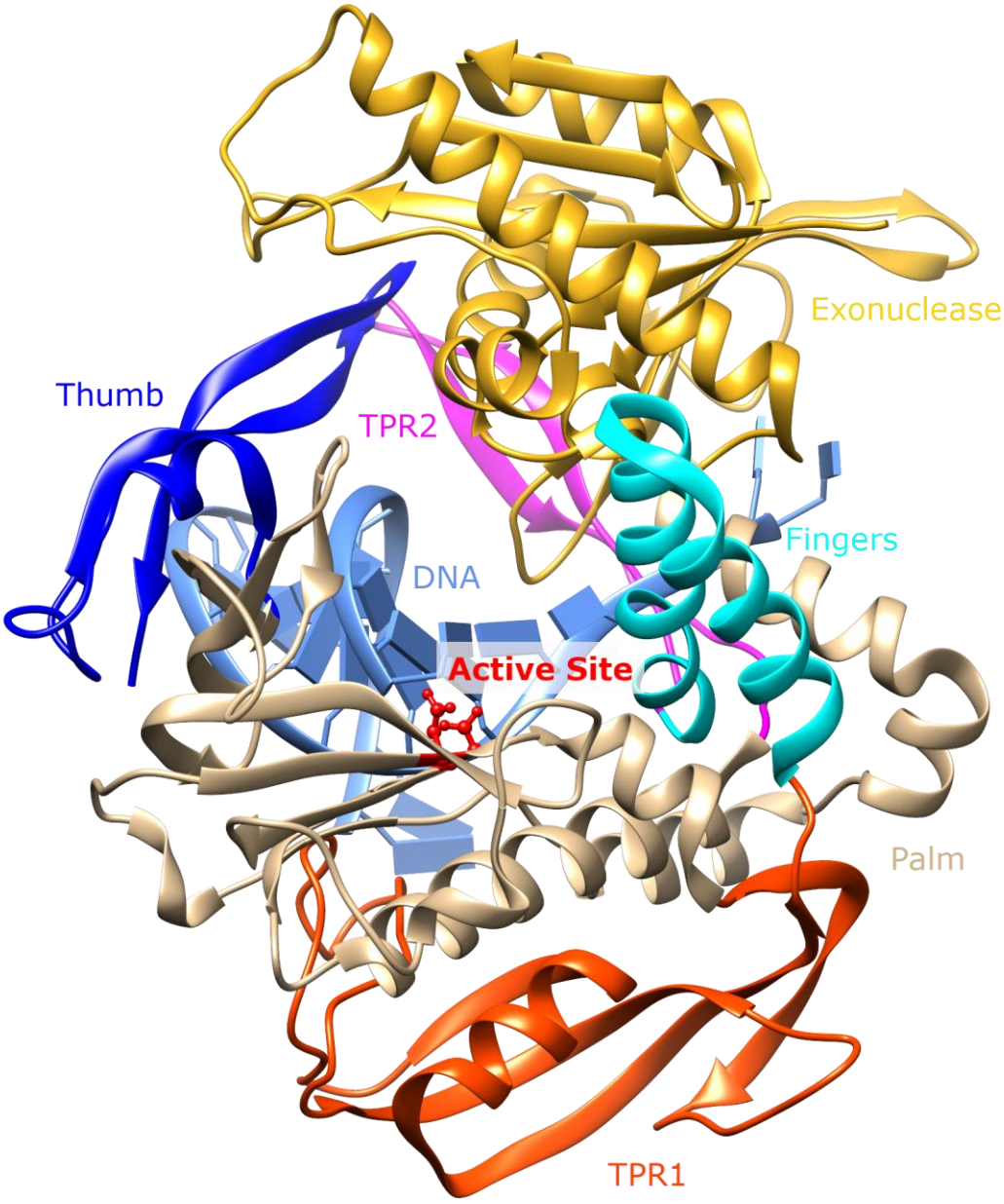

Figure 3.1: Structure of φ29 DNA polymerase, in the open conformation, bound to DNA (light blue), with each color corresponding to a separate domain. The two residues in the active site which are responsible for catalysis (D249 and D458) are near the center, highlighted in red.

Though the structure of φ29 DNA polymerase has been well-characterized, there are many open questions about the enzyme's dynamics. Recent work (114, 126) established the relative orders of nucleotide binding, pyrophosphate release, and translocation. However,

the relative orders and kinetics of the phosphodiester bond formation and the open-closed conformational changes have not been established, nor the mechanism by which the enzyme selectively closes for complementary nucleotide incorporation or whether the closing behavior exhibits any nucleotide or sequence dependence.

In this chapter, the electronic measurement method described in Chapter 1 is used to record the conformational dynamics of single φ29 DNA polymerase molecules during incorporation of nucleotides into a single-stranded DNA template. The $I(t)$ recordings show that the rate of conformational events of a single φ29 DNA polymerase molecule is strongly sequence-dependent and varies substantially in time even when processing the same template. In addition, the kinetics of the closed conformation show some nucleotide dependence and suggest the presence of an additional Poisson-like process.

3.2 Experimental Methods

3.2.1 SWCNT-FET Device Preparation

SWCNT-FET devices were fabricated according to the procedure outlined in Section 1.2 and passivated with an alumina ALD layer without PMMA. The devices were electrically characterized (procedure outlined in Section 1.3) in φ29 activity buffer (40 mM HEPES, 300 mM NaCl, 10 mM MgCl2, 100 μM TCEP, pH 6.5) solution to obtain I-V_g curves and test for RTS noise in the device. Any device SWCNT device that exhibited such RTS before polymerase attachment were excluded from further measurement and analysis.

3.2.2 φ29 DNA Polymerase Expression and Purification

An exonuclease-deficient mutant (D12A/D66A) of φ29 DNA polymerase (hereafter called φ29 DNAP) was expressed and purified by students in the laboratory of Professor Gregory Weiss (University of California, Irvine). For single-molecule measurements on SWCNT-FETs, the enzyme was dialyzed into attachment buffer (40 mM HEPES, 300 mM NaCl, 10 mM $MgCl_2$, pH 6.5) and stored at 4°C.

3.2.3 φ29 DNA Polymerase Ensemble Assays

φ29 DNAP catalytic activity was confirmed before dialysis, by students in the laboratory of Professor Gregory Weiss (University of California, Irvine), using an assay adapted from previous work (2, 127). The assay used here measured the ensemble-level polymerization of φ29 DNAP when incorporating complementary nucleotides with specific ssDNA templates: four 42 base pair (b.p.) homopolymer ssDNA templates (termed poly(dA)$_{42}$, poly(dC)$_{42}$, poly(dG)$_{42}$, and poly(dT)$_{42}$) and one 43 b.p. heteropolymer ssDNA template. Each template had an 18 b.p. M13 primer sequence on the 5'-end, and a complementary primer strand (M13F) was annealed to the primer sequence, resulting in the sequences given in Table 3.1. The positive control reactions incubated 1μM of φ29 DNAP with 5 μM of the DNA template-primer and 100 μM of each complementary nucleotide in 1x φ29 Reaction Buffer (330 mM Tris-CH_3COOH, 100 mM $MgCH_3COOH$, 660 mM KCH_3COOH, 1% (v/v) Tween 20, 10 mM DTT, pH 7.9). The negative control reactions omitted either φ29 DNAP or the

nucleotides. Reactions were kept at 30 °C for 1 h in a thermal cycler then quenched with a 2X stop solution (8 M Urea, 0.5 M EDTA, 9 mM Tris-HCl, 9 mM Boric acid) at various times.

Template	Sequence (5' to 3')
poly(dA)$_{42}$	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA<u>ACTGGCCGTCGTTTTACA</u>
poly(dC)$_{42}$	CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC<u>ACTGGCCGTCGTTTTACA</u>
poly(dG)$_{42}$	GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG<u>ACTGGCCGTCGTTTTACA</u>
poly(dT)$_{42}$	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT<u>ACTGGCCGTCGTTTTACA</u>
heteropolymer	CCTAACGCAGATAGAGACGTTGTTTAGAGATTTAAATTCGGCC<u>ACTGGCCGTCGTTTTACA</u>

Table 3.1: Sequences of the ssDNA templates, with the primer sequence underlined.

The results of the assay are displayed on agarose gels in Figure 3.2. There are five sections, corresponding to each of the four homopolymer and the one heteropolymer templates: (A) poly(dG)$_{42}$, (B) poly(dT)$_{42}$, (C) poly(dA)$_{42}$, and (D) poly(dC)$_{42}$, and (E) heteropolymer. Within each section, the leftmost column displays the results of a negative control reaction (-) run for 30 min, while the 6 columns on the right display the results of positive control reactions (+) stopped at the times indicated at the top of each column. Within each column, the presence of the ssDNA is shown as a black band near 50 b.p. (near the bottom, as indicated by the DNA ladders), and any column with polymerized dsDNA shows a black or gray band at ~80 b.p. (above the ssDNA band). Of the four homopolymer templates, only poly(dT)$_{42}$ and poly(dC)$_{42}$ showed polymerization within 30 min, and both poly(dT)$_{42}$ and the homopolymer template showed polymerization immediately (within the resolution of the measurement). These results revealed that φ29 DNAP, at the ensemble level, could completely polymerize the poly(dT)$_{42}$, poly(dC)$_{42}$, and heteropolymer templates into dsDNA but could not polymerize poly(dA)$_{42}$ or poly(dG)$_{42}$ at all.

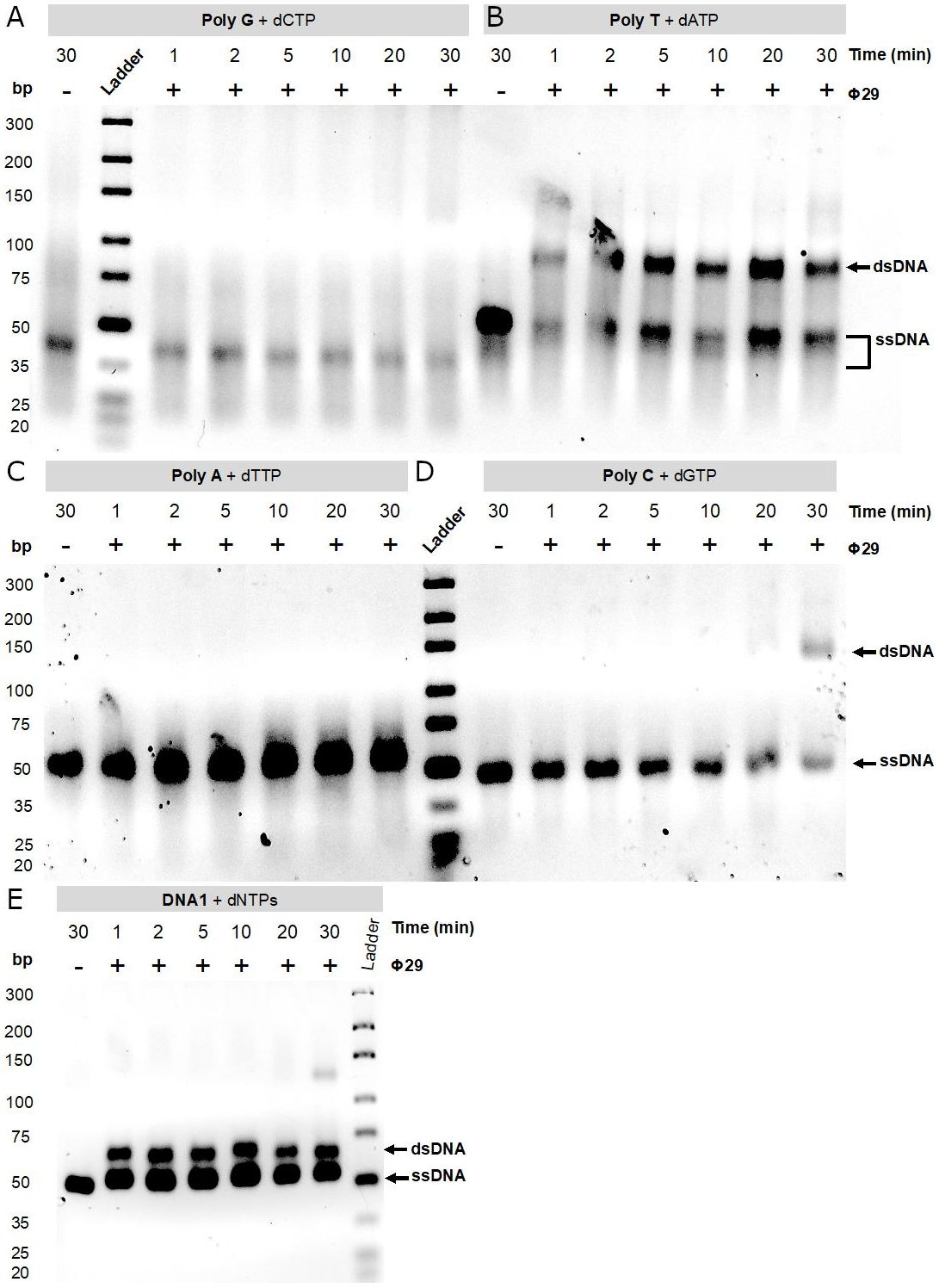

Figure 3.2: Gels showing the results of activity assays of φ29 DNAP polymerization of (A) poly(dT)$_{42}$, (B) poly(dG)$_{42}$, (C) poly(dA)$_{42}$, (D) poly(dC)$_{42}$, and (E) homopolymer. For each template, the (-) column was run without φ29 DNAP, and the remaining columns (+) were run with φ29 DNAP for the number of minutes specified at the top.

3.2.4 φ29 DNA Polymerase Attachment to SWCNT-FETs

A solution of 1 μM pyrene and 100 pM pyrene maleimide (pyrene-maleimide:pyrene ratio of 1:10,000) in dimethyl sulfoxide (DMSO) was prepared, with the pyrene acting to dilute the pyrene-maleimide on the SWCNT sidewall and limit the number of available maleimides for φ29 attachment to the SWCNT (Figure 3.3). The 1:10,000 ratio was chosen because it resulted in ~5-10 φ29 DNAP molecules attached to each SWCNT, with the φ29 DNAP molecules binding to all available maleimide molecules and saturating the number of attachment sites on the SWCNT. This number of attachments was chosen because, on average, only 1 out of every 5-10 φ29 DNAP molecules attached to the SWCNT-FET produced a signal corresponding to nucleotide incorporations, as discussed below.

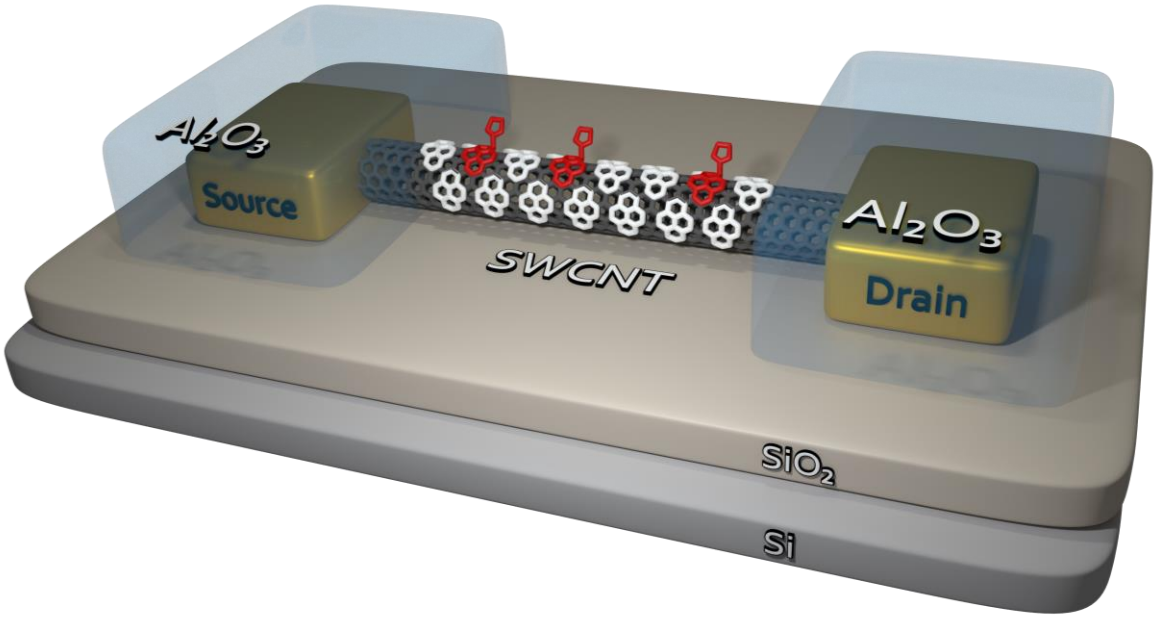

Figure 3.3: Schematic of the SWCNT-FET, with the source and drain electrodes (gold) passivated by Al_2O_3 (light blue), having both pyrene-maleimide (white) and pyrene (red) conjugated to the sidewall of the CNT (dark gray). The entire device rests on the surface of a silicon wafer (light gray) passivated by an oxide layer (SiO_2, in tan).

After cleaning the chip with RPG and rinsing with isopropyl alcohol (IPA) and then DI, the chip was submerged in the pyrene:pyrene-maleimide in DMSO solution for 2 min. Then, the chip was rinsed under a constant drip of acetonitrile for 10 s, then under flowing IPA and then DI water for another 10 s each, then finally under a constant drip φ29 activity buffer for 10 s, to remove any excess pyrene maleimide coating the CNT sidewall.

φ29 DNAP molecules were attached to the SWCNTs in solution. The φ29 DNAP, stored at 400 nM in φ29 attachment buffer at 4°C, was diluted to 4 nM with φ29 attachment buffer at 4°C, then 80 μL of the 4 nM solution was pipetted onto the surface of the chip containing selected SWCNT-FET devices. After 5 min incubation time, the chip was rinsed under a constant drip of φ29 activity buffer for 10 s to remove any excess φ29 DNAP from the chip surface, then the chip was submerged in φ29 activity buffer at room temperature (~22°C) for short-term storage. Often, the SWCNT-φ29 DNAP complex would not generate any signal within the first 12 hours after attachment, so the chip would be stored overnight in φ29 activity buffer at 4°C, and then measured again the next day.

3.2.5 Electrical Measurements of φ29 DNAP with ssDNA Templates and Nucleotides

Measurements of the SWCNT-φ29 DNAP complex with various template and nucleotide combinations were performed by immersing the complex in a solution of φ29 activity buffer containing 10 μM of ssDNA template and 100 μM of each complementary nucleotide at room temperature (~22°C). The V_g was held constant and $I(t)$ recorded for a minimum of 300 s (and often for more than 1200 s). Various values of V_g around V_T - 0.3V were tested to find

the range of V_g in which the SWCNT-φ29 DNAP would generate signal. After completing measurements for a particular template and nucleotide combination, the chip would be rinsed under a constant drip of 0.1% Tween-20 in φ29 activity buffer for 60 s, then rinsed under flowing DI for 10 s, then submerged in φ29 activity buffer until the next measurement. Between measurements of different templates, the SWCNT-φ29 DNAP complex was measured in φ29 activity buffer to ensure that previously-measured template and nucleotide did not interfere with the measurement of the next template and nucleotide.

Additional measurements were attempted with a ssDNA template comprising a 1438-b.p. sequence from green fluorescent protein (GFP), using the same procedure outlined above but with a template concentration of 100 pM. Unfortunately, even at low concentration, the GFP template proved extremely difficult to rinse away, preventing the execution of proper negative controls or subsequent measurements with different templates. Detailed analyses were not performed on the measurements from this template.

At the end of every day of measurement, the chip was stored overnight in φ29 activity buffer at 4°C. Measurements would resume the next day after allowing the chip in solution to equilibrate to room temperature for 30 min.

3.2.6 Atomic Force Microscopy Imaging

Atomic force microscopy (AFM) (Pacific Nanotechnology Nano-R) was used to determine how many φ29 DNAP molecules were attached to the SWCNT (Figure 3.4), after all $I(t)$

measurements were completed. Attached φ29 DNAP appeared as dots with heights ~3 nm overlapping with the line of the SWCNT.

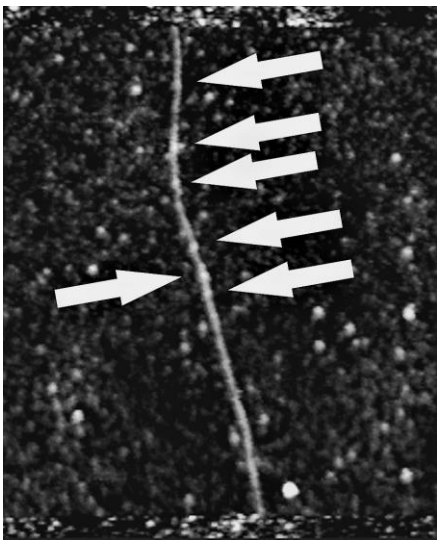

Figure 3.4: AFM image of an example SWCNT-φ29 DNAP complex. The individual φ29 DNAP molecules appear as dots (shown with an arrow) overlapping with the line of the SWCNT, here shown running down the middle of the image.

3.2.7 Signal Processing and Analysis

The raw $I(t)$ signal was converted to a simple binary signal to facilitate two-state analysis, as discussed in Section 1.5. Each data point in the detrended signal was assigned to either a high state (baseline) or low state (current excursion) based on a simple threshold algorithm, with the threshold placed approximately halfway between the baseline and the highest-magnitude current excursion, and with the algorithm set to ignore events shorter than 80 μs. The duration for each occurrence of the high and low states was calculated, and the durations of each state was collected into probability distributions and displayed on semi-log plots.

In addition, the entire >300 s data set was analyzed to calculate the rate of current excursions and to calculate the percentage of time that the φ29 DNAP molecules were active. After detrending and assigning data points to the high or low state, the number of excursions per second was calculated for every second of the data set.

Due to the multi-molecule attachment scheme, two or more φ29 DNAP would occasionally exhibit conformational activity during the same period, producing multiple levels (3 or more) in the $I(t)$. Devices exhibiting such behavior were not analyzed for single-molecule kinetics, but the $I(t)$ recordings could still be studied for nucleotide-specific characteristics.

3.3 Results

SWCNT-φ29 DNAP complexes generated signals that can be described by three categories. Some complexes produced no $\Delta I(t)$ signal above baseline under any condition (Figure 3.5A), while others produced constant fluctuations in $\Delta I(t)$ despite the absence of DNA template and nucleotides (Figure 3.5B), and others produced template-dependent excursions in $\Delta I(t)$ (Figure 3.5C) when exposed to DNA template and complementary nucleotides but did not produce signal when the template and nucleotide were washed away. Such excursions were like those seen in previous experiments for the Klenow Fragment of DNA polymerase I (2, 127), and indicated that the observed signal is correlated with nucleotide incorporation events. The remainder of this study focuses on the SWCNT-φ29 DNAP complexes which produced template-dependent signal like the example shown in Figure 3.5C.

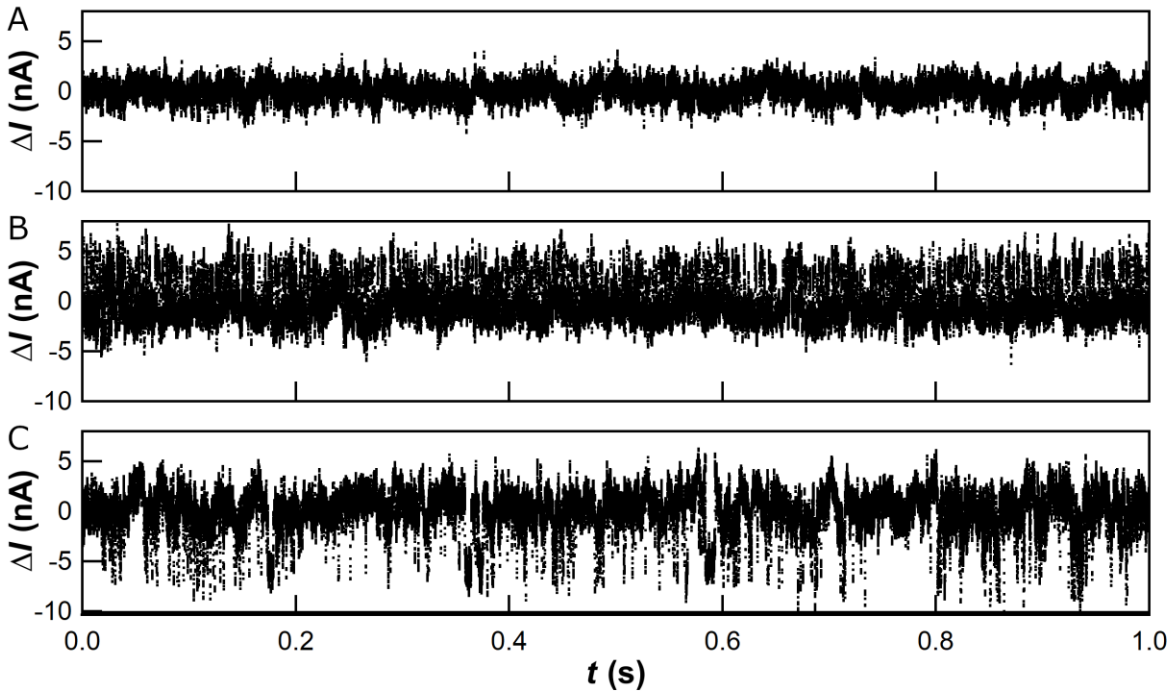

Figure 3.5: Example $\Delta I(t)$ for SWCNT-φ29 DNAP complexes which: (A) produced no current excursions, (B) always produced excursions, or (C) produced excursions only when exposed to template and complementary nucleotide.

The SWCNT-φ29 DNAP complexes within each signal-producing category exhibited signal within specific ranges of V_g. Figure 3.6 shows normalized I-V_g curves (thin lines) and the range of signal activity (thick lines) for the complexes which always produced current excursions (A), which showed signal for $V_g - V_T \leq -0.5$ V, and the complexes producing template-dependent excursions (B), which were active for -0.5 V $\leq V_g - V_T \leq -0.2$ V. The value $V_g - V_T$ is a more reliable indicator (than V_g alone) of the effective gating experienced by the SWCNT-φ29 DNAP complex, since shifts in V_T reflect the electrostatic environment experienced by the SWCNT and compensate for environmental shifts in V_g. For each device in the template-dependent category of SWCNT-φ29 DNAP complexes, current excursions were most frequent when the V_g was held within 25 mV of the center of the active range, though current excursions persisted at V_g up to 50 mV from the center.

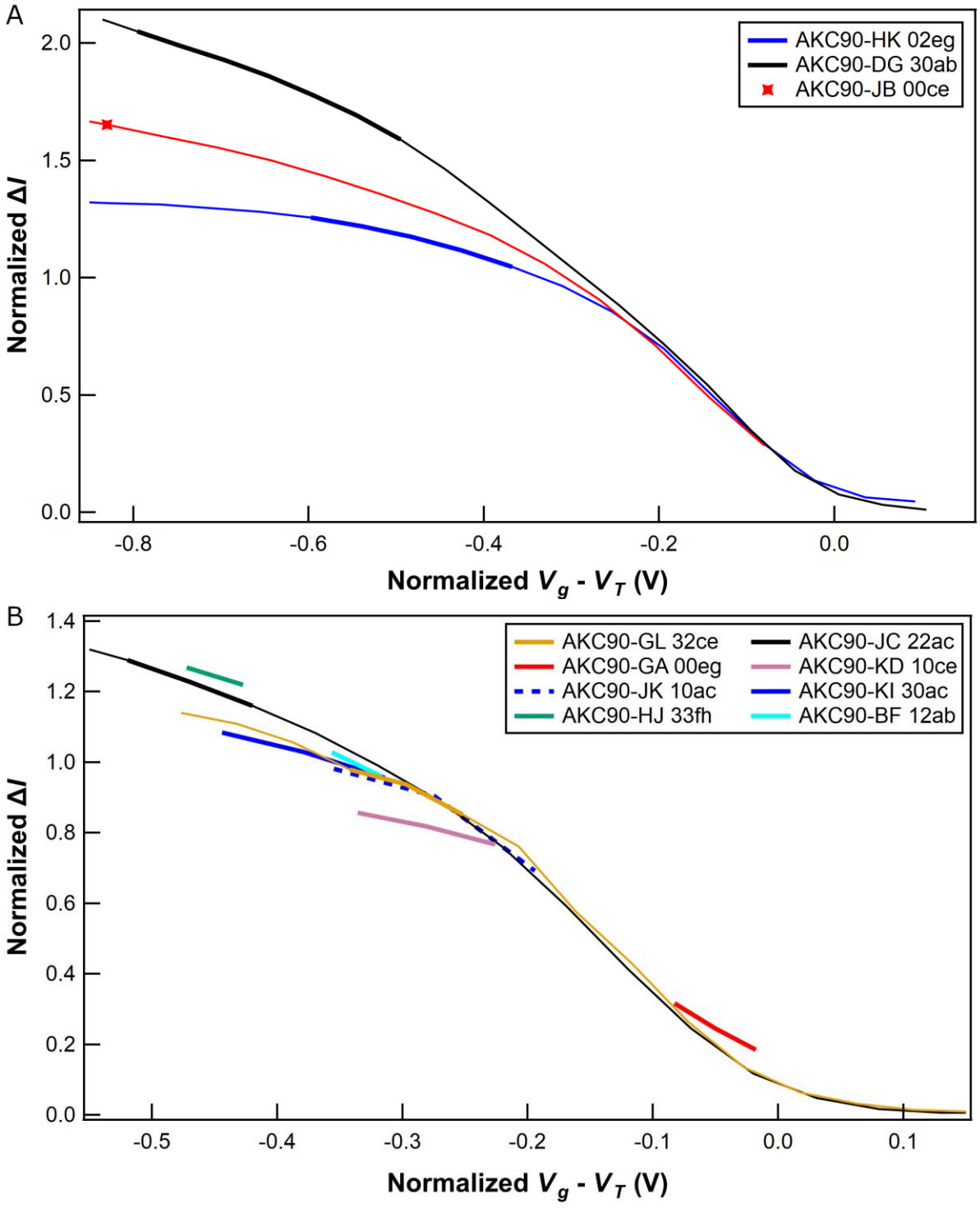

Figure 3.6: Normalized I-V_g curves highlighting the V_g range (thick lines) relative to V_T (0 V) over which current excursions were observed for SWCNT-φ29 DNAP complexes which: (A) always produced excursions, and (B) produced template-dependent excursions. The thin lines show the full I-V_g curves for several devices for comparison.

For each measurement with DNA template and complementary nucleotide, the $\Delta I(t)$ resided mostly at the baseline level ($\Delta I(t) = 0$), with brief downward excursions below the baseline to a lower level. Figure 3.7 shows representative $\Delta I(t)$ signals (left) and the corresponding histograms (right) produced by a single SWCNT-φ29 DNAP complex when the φ29 DNAP is active. The top pair of graphs correspond to the $\Delta I(t)$ from a SWCNT-φ29 DNAP complex in only φ29 activity buffer, and the bottom five pairs of graphs show the signal when the SWCNT-φ29 DNAP complex is exposed to the indicated template and complementary nucleotides. The $\Delta I(t)$ from the buffer measurement (Figure 3.7A) showed no current excursions throughout the measurement. The $\Delta I(t)$ from the heteropolymer template (Figure 3.7B) exhibited a moderate rate of sporadic downward excursions, sometimes occurring in short bursts, with a mixture of both short- and long-duration excursions, each with approximately similar amplitudes from baseline to peak. The long excursions were infrequent, but some were of similar duration to the waiting time between excursions. The $\Delta I(t)$ from poly(dT)$_{42}$ (Figure 3.7C) contained a sporadic mixture of both short and long downward excursions. The $\Delta I(t)$ from poly(dA)$_{42}$ (Figure 3.7D) and poly(dG)$_{42}$ (Figure 3.7E) contained bursts of short-duration downward excursions, even though the activity assay (Figure 3.2A and C) showed no catalytic activity. The baseline of the $\Delta I(t)$ from poly(dG)$_{42}$ exhibited larger fluctuations in comparison to the other templates, resulting in wider baseline and secondary peaks in the histogram. The $\Delta I(t)$ from poly(dC)$_{42}$ (Figure 3.7F) contained periods of a rapid rate of short downward excursions spaced closely together, with few pauses or quiet periods.

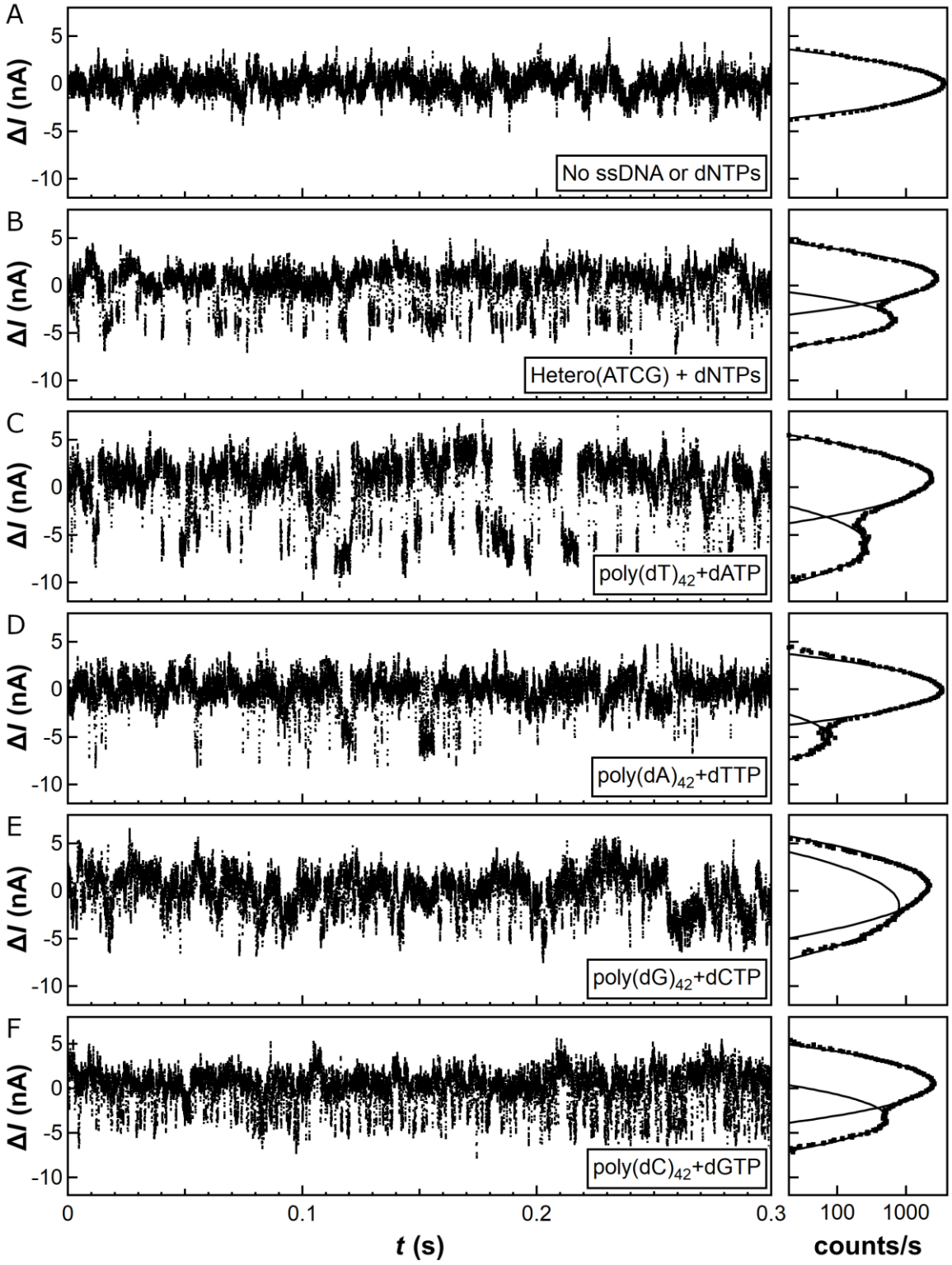

Figure 3.7: Example $\Delta I(t)$ signals (left) and histograms of the $\Delta I(t)$ (right) for: (A) buffer-only measurement, followed by measurements with the (B) heteropolymer template and (C-F) four homopolymer templates.

Each graph on the right of Figure 3.7 displays a histogram of the corresponding $\Delta I(t)$. The bottom five histograms contain a second peak corresponding to the downward excursions. The histogram for the $\Delta I(t)$ of the buffer-only measurement (top) lacks the second Gaussian peak, reflecting the lack of downward excursions in the $\Delta I(t)$. The second peak for poly(dA)$_{42}$ data was much lower in amplitude than the second peak for any other template due to the low frequency and short duration of the current excursions. Each peak in the histogram was fit to a Gaussian function, shown as black lines in the histogram. The Gaussian fits indicate the positions of the two current states, along with the width of each state. The $\Delta I_{h\text{-}l}$, which is the amplitude of the $\Delta I(t)$ between the high and low states according to the fitted Gaussian, was calculated for each template. Overall, the $\Delta I_{h\text{-}l}$ was highest for poly(dT)$_{42}$, then poly(dA)$_{42}$, then poly(dC)$_{42}$ and the heteropolymer templates with about the same amplitude, and then lowest for poly(dG)$_{42}$. In addition, the width of the baseline was higher for poly(dG)$_{42}$ than for any of the other templates due to the increased baseline noise.

Previous studies with the Klenow Fragment of DNA polymerase I (2, 127) demonstrated that such downward excursions from the high state to the low state resulted from the enzyme's mechanical conformational change from the open to the closed conformation, triggered by the binding of the correct nucleotide to the fingers region of the polymerase. Similarly, the downward excursions observed from φ29 DNAP were assigned to the low state, corresponding to the closed conformation of φ29 DNAP, and the baseline assigned to the high state, or open conformation.

The kinetics of φ29 DNAP conformational activity for each of the DNA templates was quantified using two methods. In the first method, the calculated durations of the open and closed states from each dataset were collected into separate histograms, then normalized to form probability distributions. Figure 3.8 shows, on a log-linear plot, the probability distributions for the waiting time between downward excursions for each of the measured DNA templates, calculated as the probability per 1 ms bin. Each waiting time distribution approximates a stretched exponential or double-exponential function, with the shortest durations being the most probable. All the distributions overlap except the one for poly(dA)$_{42}$, which has a distinctively shallower slope than the other distributions. This indicates that φ29 DNAP has a higher probability for long waiting times between events when processing poly(dA)$_{42}$ than when processing other DNA templates.

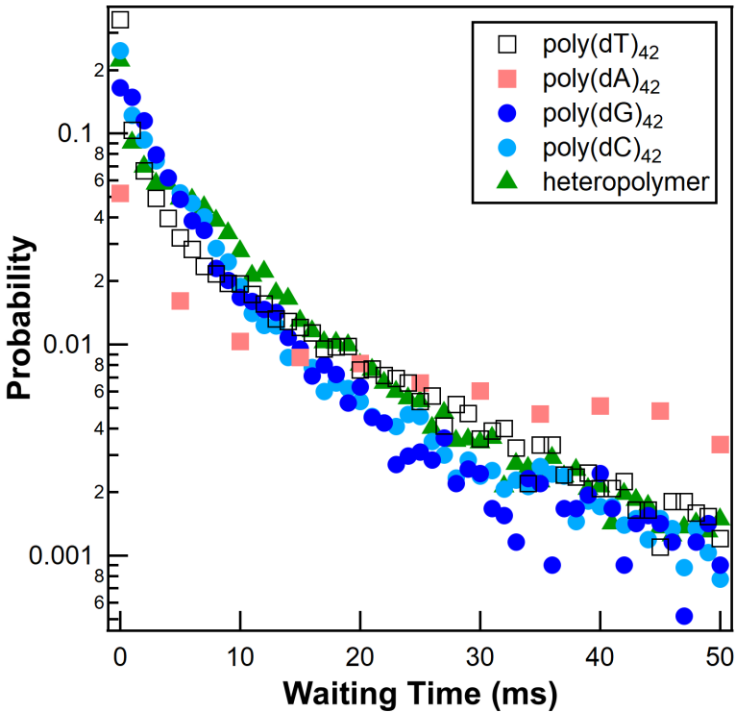

Figure 3.8: Probability distributions of waiting times between events for each template. Most of the distributions overlap, excepting the distribution for poly(dA)$_{42}$.

In the second method to quantify the kinetics of φ29 DNAP conformational activity, the rate of downward current excursions was calculated for the entire dataset for each of the DNA templates to evaluate the time dependence of the conformational event rate. Figure 3.9 shows, on the left, plots of the rate per second of excursions over time for all five DNA templates, and on the right, the histograms of the rate (plotted vertically) on a log-linear scale. Each plot shows the rate over one continuous measurement, with the duration of each measurement (in seconds) shown on the bottom axis. Any rates below 10 s^{-1} were indistinguishable from the noise background due to the inherent noise from the SWCNT-FET, so the enzyme was considered inactive (not exhibiting conformational motion) during periods of such rates. Generally, the rate during active periods varied stochastically from one second to the next, with brief bursts of rapid rates (> 100 s^{-1}) interspersed among longer periods of moderate (~ 25-100 s^{-1}) or even low (< 25 s^{-1}) rates. For poly(dA)$_{42}$ (B) and poly(dG)$_{42}$ (C), the downward excursions occurred in short bursts, at most 2 s long and not correlated with the length of the template. Periods of little or no activity lasted up to 400 s, and the histograms show that the rate remained between 0-20 s^{-1} for >90% of the time. For the remaining templates, the polymerase exhibited continuous activity during discrete periods of up to ~1000 s. The rate fluctuated the most for poly(dC)$_{42}$ (D), exhibiting jumps between discrete levels of rates, ranging between 50 s^{-1} and 400 s^{-1}, and spending <100 s at each level before jumping to the next. The histograms of the rate contain slight bumps, highlighting rates that were more common during the measurement. For instance, the histogram for poly(dT)$_{42}$ (A) has local maxima at ~20 s^{-1} and ~100 s^{-1}, while the histogram for poly(dC)$_{42}$ (D) has a broad hump at ~250 s^{-1} in addition to the peak at 0 s^{-1}.

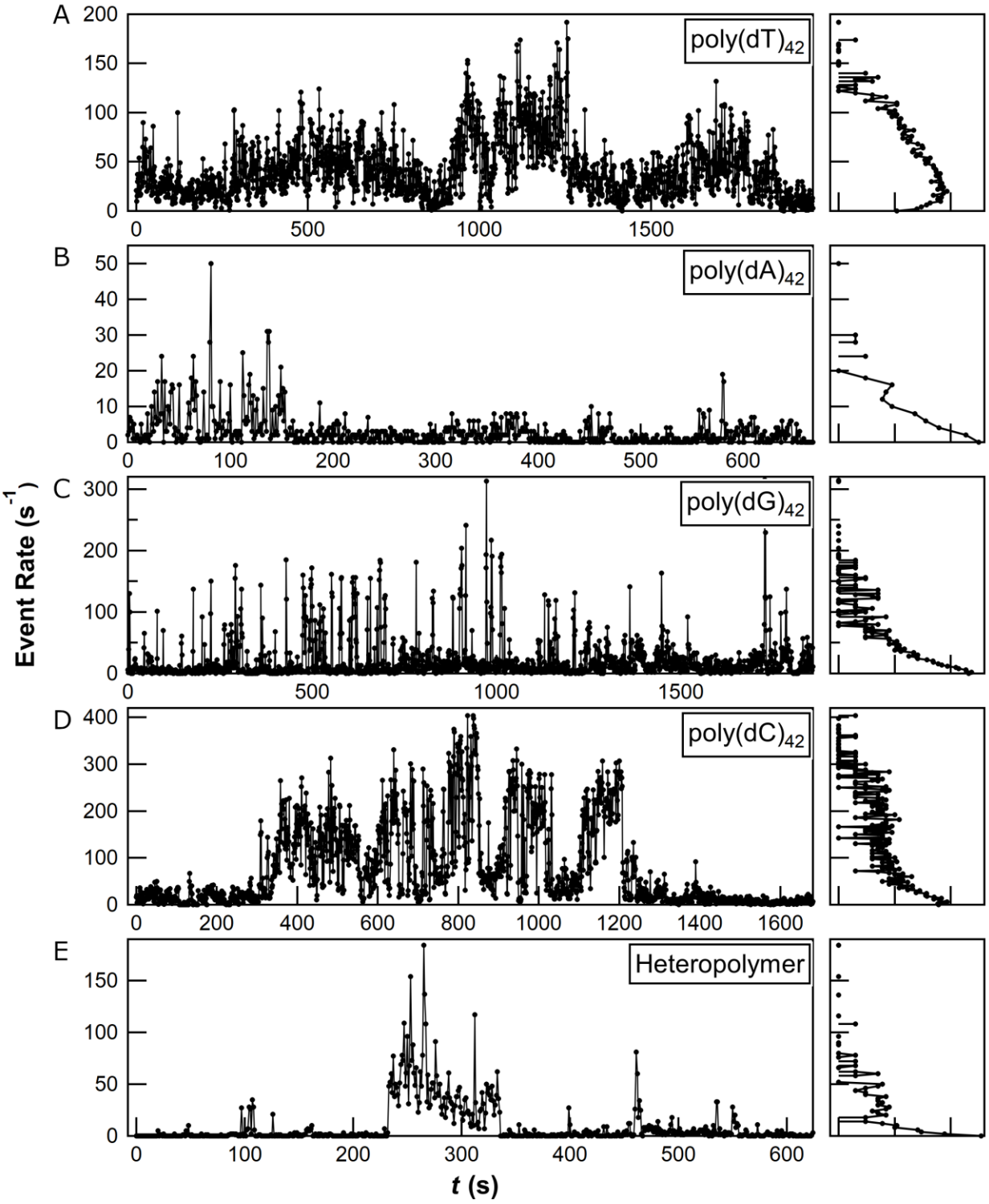

Figure 3.9: Plots of the event rate (left) and histograms of the event rate (right) for the (A-D) four homopolymer templates and (E) heteropolymer template.

Further analysis of periods of high conformational activity reveal additional dynamics in the conformational event rate. Figure 3.10A and B zooms in on a 180 s active period in both the poly(dT)$_{42}$ and poly(dC)$_{42}$ measurements, showing the rate over time in the left graph, and the histogram of the rate (on a linear scale) in the right graph. The rate for poly(dT)$_{42}$ (Figure 3.10A) fluctuated relatively slowly, being approximately constant during this active period and showing small fluctuations as expected from a stochastic Poisson process. By contrast, the event rate for poly(dC)$_{42}$ (Figure 3.10B) exhibited sharp jumps between "fast" and "slow" rates, spending 2-50 s at one rate before switching to the other, and the φ29 DNAP switched stochastically between these two rates throughout the entire active period. A histogram of the rate revealed two distinct Gaussian peaks corresponding to two distinct processing modes, with rates at ~40 s^{-1} (slow) and ~190 s^{-1} (fast). Figure 3.10C and D show examples of the $I(t)$ corresponding to the fast (C) and slow (D) rates. In Figure 3.10C, the density of current excursion events is high and the waiting time between events is small (~5 ms), whereas in Figure 3.10D, the density of current excursions is lower and the waiting times between events is longer (~20 ms). These examples show that the enzyme follows one rate for periods of several seconds before switching to the other rate, and that the conformational motions within each period are consistent. The distribution of rates and the consistency of events corresponding to each rate suggests that, at least for processing poly(dC)$_{42}$, the enzyme switches between two processes, adopting one rate for a few seconds before switching to the other rate.

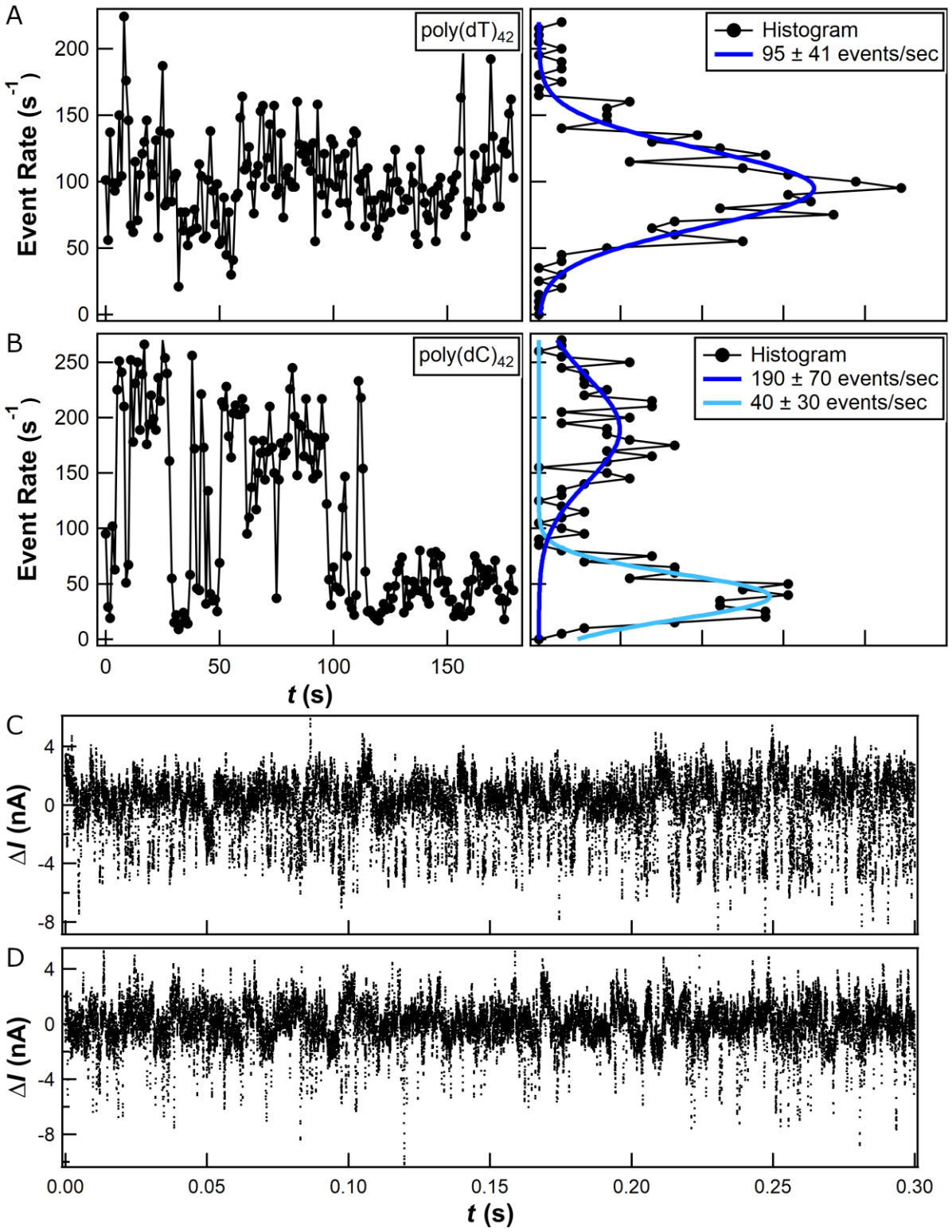

Figure 3.10: (A and B) Event rates (left) and histograms of the rate (right) for a 180 s period of φ29 DNAP when processing (A) poly(dT)$_{42}$ and (B) poly(dC)$_{42}$. The histogram for poly(dC)$_{42}$ shows two peaks, indicating two discrete rates. (C and D) Examples of the $\Delta I(t)$ corresponding to the faster (C) and slower (D) rates from poly(dC)$_{42}$.

The single-molecule $I(t)$ recordings of poly(dT)$_{42}$ revealed a proportion of current excursions with durations much longer than expected from a single Poisson process. Figure 3.11A shows a 0.5 s example of the $\Delta I(t)$ from poly(dT)$_{42}$, showing many excursions that are ~100-200 μs in duration but also showing several that are >5 ms in duration. These longer excursions were seen in the $I(t)$ for measurements of templates containing thymine bases, such as poly(dT)$_{42}$, the heteropolymer template, and the GFP template, but not for templates lacking thymine. For comparison, the excursions in $\Delta I(t)$ from the remaining homopolymer templates (displayed in Figure 3.11B) exhibit durations ~ 100 μs and all appear similar.

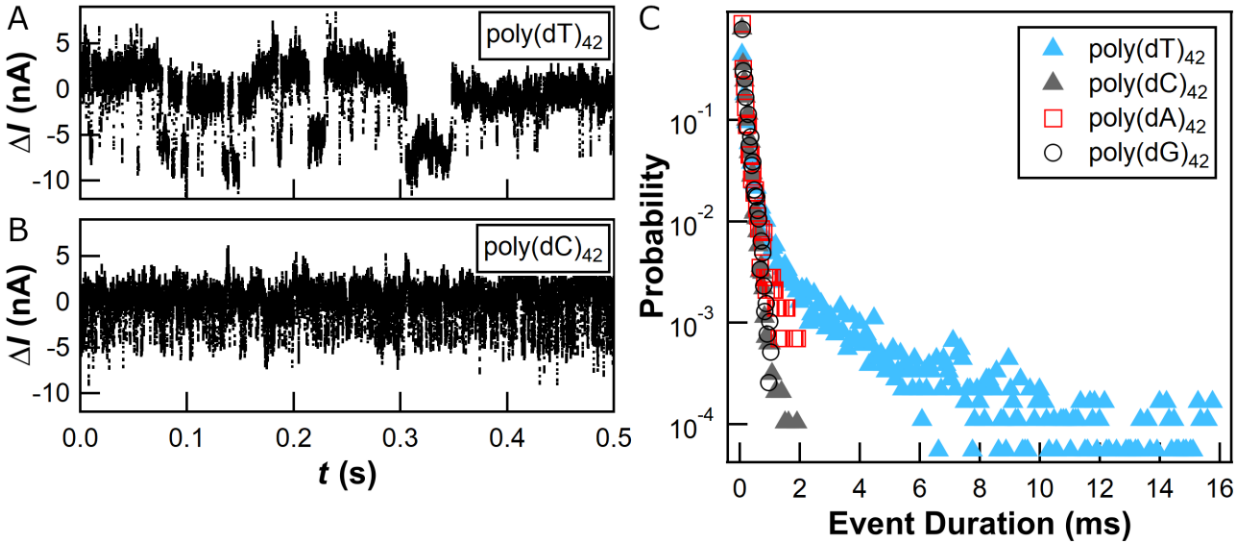

Figure 3.11: Examples $\Delta I(t)$ from measurements of (A) poly(dC)$_{42}$ and (B) poly(dT)$_{42}$, showing the presence of long-duration closed events with poly(dT)$_{42}$. (C) Probability distributions of the duration of closed events for poly(dC)$_{42}$ and poly(dT)$_{42}$, where the long-duration closed events appear as a second exponential beyond 2 ms.

To quantify the likelihood of these long-duration excursions, Figure 3.11C shows the probability distributions for the closed state durations for all four homopolymer templates. Each distribution is normalized to the probability per 80 μs bin. The poly(dC)$_{42}$, poly(dA)$_{42}$,

and poly(dG)$_{42}$ distributions overlap neatly on a straight line, corresponding to a single exponential function. The poly(dT)$_{42}$ distribution overlaps the other distributions until ~1 ms, at which point it diverges to a much shallower slope. The single exponential form of the non-poly(dT)$_{42}$ distributions suggest that the duration of the closed state is governed by a single Poisson process, with characteristic time τ = 130 μs (except for poly(dA)$_{42}$, which exhibited a slightly longer time constant). By contrast, the double-exponential distribution for poly(dT)$_{42}$ suggests that the closed state alternates between two different Poisson processes with characteristic times τ_1 = 130 μs and τ_2 = 1 ms.

The characteristic rates and times for each template, along with the amplitudes between the open and closed states, are listed in Table 3.2.

Template	Avg. Rate (s^{-1})	Rate1 (s^{-1})	Rate2 (s^{-1})	Closed τ_1 (μs)	Closed τ_2 (μs)	Open τ_1 (ms)	Open τ_2 (ms)	ΔV_g (mV)	σ_{prim} (mV)	σ_{sec} (mV)
poly(dT)$_{42}$	21	100±40	-	170±90	1440±700	13±7	-	44	13	16
poly(dA)$_{42}$	3	-	-	220±110	-	30±15	-	32	11	13
poly(dG)$_{42}$	21	-	-	130±70	-	7±4	-	12	16	19
poly(dC)$_{42}$	80	40±30	190±70	130±70	-	5±3	30±15	24	13	13
Hetero	10	60±40	140±60	130±70	-	8±4	40±20	27	11	10

Table 3.2: Processing rates, characteristic times τ, and signal amplitudes exhibited by φ29 DNAP while processing various templates and complementary nucleotides.

3.4 Discussion

In this study, single-molecule attachments were initially attempted for φ29 DNAP. However, most single-molecule attachments failed to produce any $\Delta I(t)$ excursions above baseline. In fact, even with an average of 5 attached φ29 DNAP molecules per SWCNT, only one out of

every two SWCNT-φ29 DNAP devices produced any template-dependent $\Delta I(t)$ excursions, resulting in an effective yield of about 1 attachment with signal for every 10 attachments. The remaining φ29 DNAP attached to the SWCNT sidewall were either conformationally inactive or attached in an orientation that precluded signal transduction to the SWCNT-FET.

Previous studies (2, 127) of the Klenow Fragment of DNA polymerase I examined a single DNA polymerase enzyme attached to a single SWCNT-FET, resulting in approximately 1 clear single-molecule observation for every 2 devices. Those studies utilized a single-cysteine mutant of the Klenow Fragment that facilitated a consistent attachment location when linking the polymerase to the SWCNT-FET with the maleimide-thiol reaction. Thus, any attached Klenow Fragment molecule always presented the same residues to the SWCNT-FET, so any signal produced by one SWCNT-Klenow complex were directly comparable to signal from another complex. Unfortunately, native φ29 DNAP possesses seven cysteines, four of which were shown (using matrix-assisted laser desorption/ionization, or MALDI) to attach to maleimide groups in solution. The four maleimide-binding cysteines are highlighted in blue in Figure 3.12, and the three non-maleimide-binding cysteines are colored red. Multiple attempts to produce single-cysteine or even reduced-cysteine mutants resulted in either no expression or inactive enzyme, so all measurements were conducted with the "wild-type" φ29 DNAP containing all seven cysteines.

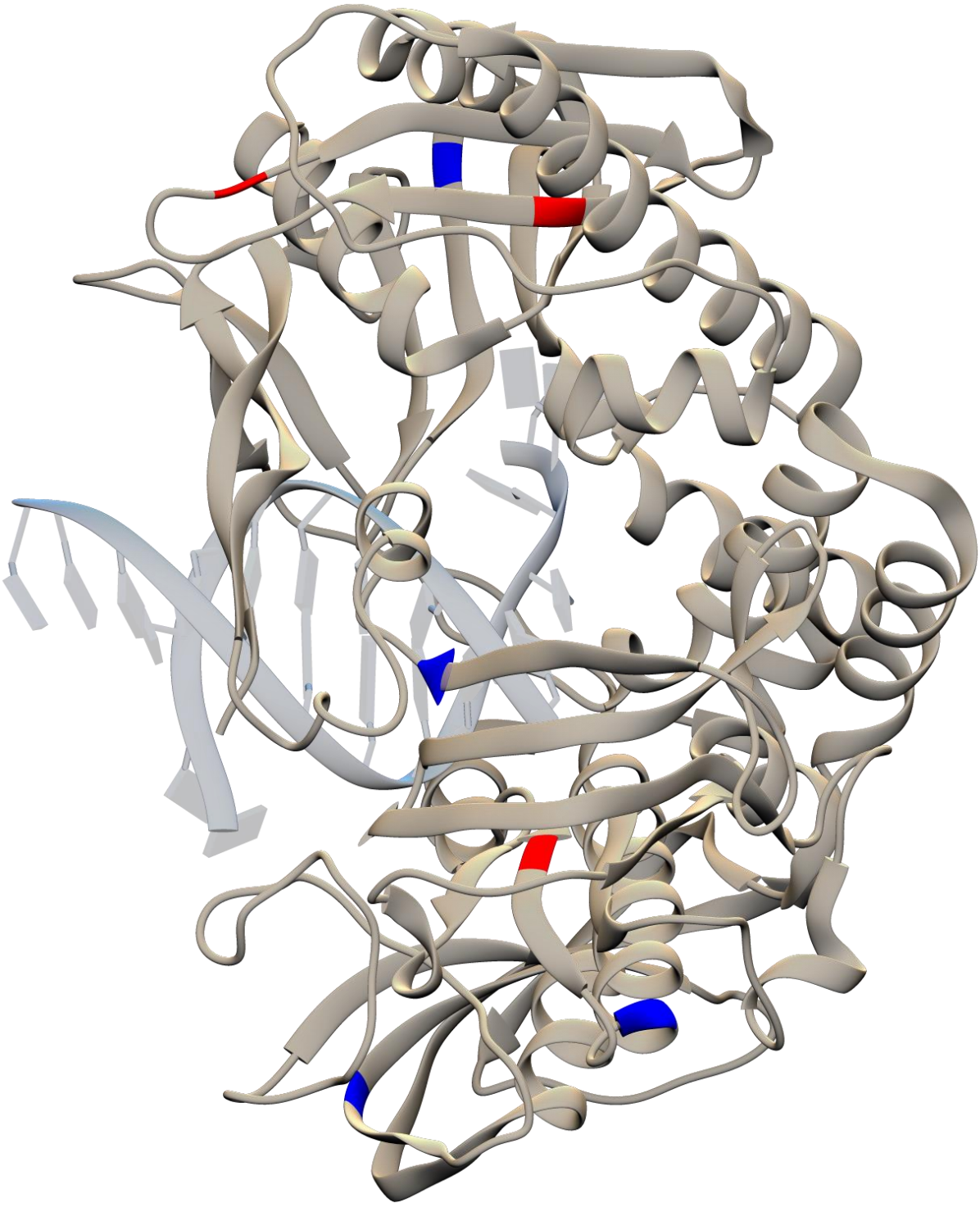

Figure 3.12: Structure of wild-type φ29 DNAP, with the enzyme in tan and the DNA in light gray. The seven cysteines in the enzyme are highlighted in color: blue for the four cysteines that react with maleimides in solution, and red for the three that do not.

Due to the presence of multiple exposed cysteines, every φ29 DNAP attached to the SWCNT sidewall resided in one of four possible orientations, with each orientation presenting a different set of residues to the SWCNT. Thus, the resulting signal from each attached φ29 DNAP depends on its orientation, with some orientations producing no signal at all (such as in Figure 3.5A), other orientations producing constant current excursions (either up or down, with an example shown in Figure 3.5B), and still others producing the template-dependent signal already discussed (Figure 3.5C). The statistics of this experiment already show that the most productive orientation (which produces template-dependent signal) is statistically unfavored during maleimide conjugation. Further tests, perhaps using MALDI as mentioned above, might be able to statistically determine which cysteines are most likely to bind with the maleimide groups. Then, the most productive orientation could be established by developing a procedure to block all but one of the cysteines so that each orientation can be repeatedly tested in a controlled manner.

The overlap among the probability distributions of the waiting times between conformational events for most of the DNA templates (Figure 3.8) showed that the typical waiting time was similar, despite the different event rates observed in the $\Delta I(t)$. This suggests that the difference in rates was not due to the "typical" short waiting time, as shown in Figure 3.8, but rather to the presence and quantity of the "atypical" long waiting times, which appeared as pauses longer than 200 ms. The unusually shallow slope of the poly(dA)$_{42}$ distribution likely reflects the extreme difficulty encountered by φ29 DNAP when attempting to process that template. More generally, the stretched exponential form of the waiting time probability distributions suggests that the waiting times were not the result of

a single Poisson process, but could be due a combination of several independent processes or a process of several correlated steps. Previous studies (105, 128) established that the closing transition was not the rate-limiting step for DNA polymerase I and Taq DNA polymerase, and this could be the same for φ29 DNAP. Some possible reasons for the stretched exponential distribution could be: pauses due to snags when unraveling the secondary structure in the ssDNA template, delays in nucleotide binding or in releasing the pyrophosphate, or dynamic disorder driven by spontaneous thermal fluctuations that slightly change the effective processing rate over time.

The highly variable nature of the event rate shows that the rate of φ29 DNAP conformational motion remained stochastic even to the timescale of seconds and minutes. φ29 DNAP can pause for long periods of time, as observed in previous studies (111, 112), but can also exhibit rapid conformational changes in short bursts of up to ~60 s long. This study shows that pauses were template-dependent, with activity on poly(dA)$_{42}$ or poly(dG)$_{42}$ exhibiting substantially longer pauses than activity on poly(dT)$_{42}$ or poly(dC)$_{42}$.

In addition, the average rate, which is the metric most comparable to ensemble measurements, varied drastically with template and showed that some conformational motions are non-catalytic and do not correspond to nucleotide incorporations. For example, recordings of poly(dC)$_{42}$ and poly(dT)$_{42}$ exhibited average conformational event rates of ~80 s^{-1} and ~20 s^{-1}, respectively. By contrast, the ensemble activity assay showed greater catalytic activity for poly(dT)$_{42}$ than poly(dC)$_{42}$, indicating that φ29 DNAP exhibited more catalytic conformational motion when processing poly(dT)$_{42}$ than poly(dC)$_{42}$. Thus, the

large number of conformational events in the poly(dC)$_{42}$ recordings must contain a mixture of catalytic and non-catalytic motions. In particular, bursts of conformational motion with rates exceeding 200 s^{-1}, which occurred for ~40% of the active periods in the poly(dC)$_{42}$ recording, were likely dominated by non-catalytic motions.

The $\Delta I(t)$ for poly(dA)$_{42}$ or poly(dG)$_{42}$ showed some, though infrequent, downward excursions, which contrasted with the result from activity assays used in this study. The ensemble activity assays of φ29 DNAP suggested that the polymerase does not process poly(dA)$_{42}$ or poly(dG)$_{42}$ at all, but the single molecule $I(t)$ showed that the polymerase still exhibited some conformational motion in its attempts to process the templates. This might suggest that φ29 DNAP struggles to process these templates but still exhibits a small amount of catalytic activity. Other studies (112, 129, 130) showed that templates with repeating bases or patterns, including GC repeats, possess secondary structure that hinder DNA replication. Specifically, templates with repeating guanine bases can form G-quadruplex or other structures (131, 132), which polymerases find difficult to unravel. In addition, poly(dA) is known to form straight, rod-like helices in solution that are more rigid than those for poly(dT) (133-136), and this rigidity may prevent the template from threading into the active site in φ29 DNAP. Thus, the burst-like nature and the frequent pauses in φ29 DNAP activity when processing poly(dA)$_{42}$ or poly(dG)$_{42}$ could be explained by the difficulty encountered by φ29 DNAP in unraveling the secondary structure of specific templates.

The second process seen when φ29 DNAP processes templates containing thymine, which resulted in a long-duration closed event, is not understood. These long-duration events are

not directly correlated with nucleotide incorporations, since they occur at rates much lower than the known ensemble catalytic rate, and do not appear in recordings of activity on poly(dC)$_{42}$ even though the template is catalyzed at the ensemble level. Furthermore, the long-duration events are an order of magnitude longer than events corresponding to catalysis in the Klenow Fragment of *E. coli* DNA polymerase I (2). In addition, these long-duration events are unlikely to be due to an error-checking mechanism, since the φ29 DNAP used here has a disabled exonuclease domain and cannot perform error-checking. One possibility is that the long-duration closed events occur when the φ29 DNAP jams during translocation of the template to the position of the next base. Another possibility is that the incorporation is delayed because the bases are oxidized or otherwise damaged, either in the template or in the nucleotides in solution, which might impede the closed-open conformational change or require extra steps for the bond to complete.

3.5 Summary

Single-molecule investigations of φ29 DNA polymerase revealed that the enzyme's efficiency in processing ssDNA was highly dependent on template composition. The enzyme readily processed templates containing mostly thymine or cytosine bases, but exhibited pauses and short bursts of conformational motion when processing templates containing long stretches of adenine or guanine bases. The single-molecule SWCNT-FET technique was able to observe the sparse conformational activity of φ29 DNA polymerase on homopolymer templates containing adenine (A) and guanine (G) bases even when the ensemble assays failed to detect catalytic activity. Analysis of the enzyme's processing rates showed that the rates varied

with time, sometimes even alternating between two different modes. In addition, the distribution of the durations of the closed conformation demonstrated the existence of a second Poisson-like process when the enzyme processed templates containing thymine (T) bases. Further studies are needed to determine the mechanisms behind the alternating processing rates or additional processes that were uncovered in this work.

CHAPTER 4

Noise Reduction and Signal Processing Methods

4.1 Introduction

DNA polymerases and other biomolecules generate complex signals in SWCNT-FET recordings, exhibiting sharp transitions and spikes over a large range of timescales spanning the ms and μs range. Unfortunately, the SWCNT-FET sensor itself produces significant noise in similar timescales, reducing the signal-to-noise (SNR) ratio in single-molecule measurements and even obscuring individual conformational events when the SNR ratio drops below 2:1. Biochemical approaches to increasing the SNR, such as using mutagenesis to increase the effective gating experienced by the SWCNT-FET during a conformational event (25), involve long lead times and low odds for producing a catalytically-active enzyme, as encountered both for φ29 DNA polymerase and the Klenow Fragment of *E. coli* DNA polymerase I. In addition, increasing the measurement bandwidth to acquire more information in the raw $I(t)$ signal allows more noise into the acquired signal and results in a larger peak-to-peak noise amplitude in the signal baseline. Figure 4.1A shows an example $\Delta I(t)$ from the 25 kHz probe station (as described in Section 1.3), which has a ~5 nA peak-to-peak baseline noise amplitude, while Figure 4.1B shows an example $\Delta I(t)$ from the 220 kHz flow cell, which has a ~10 nA peak-to-peak baseline noise amplitude. Thus, some type of post-acquisition digital noise reduction is necessary to reduce the baseline noise amplitude while preserving the sharp spikes corresponding to biomolecule activity.

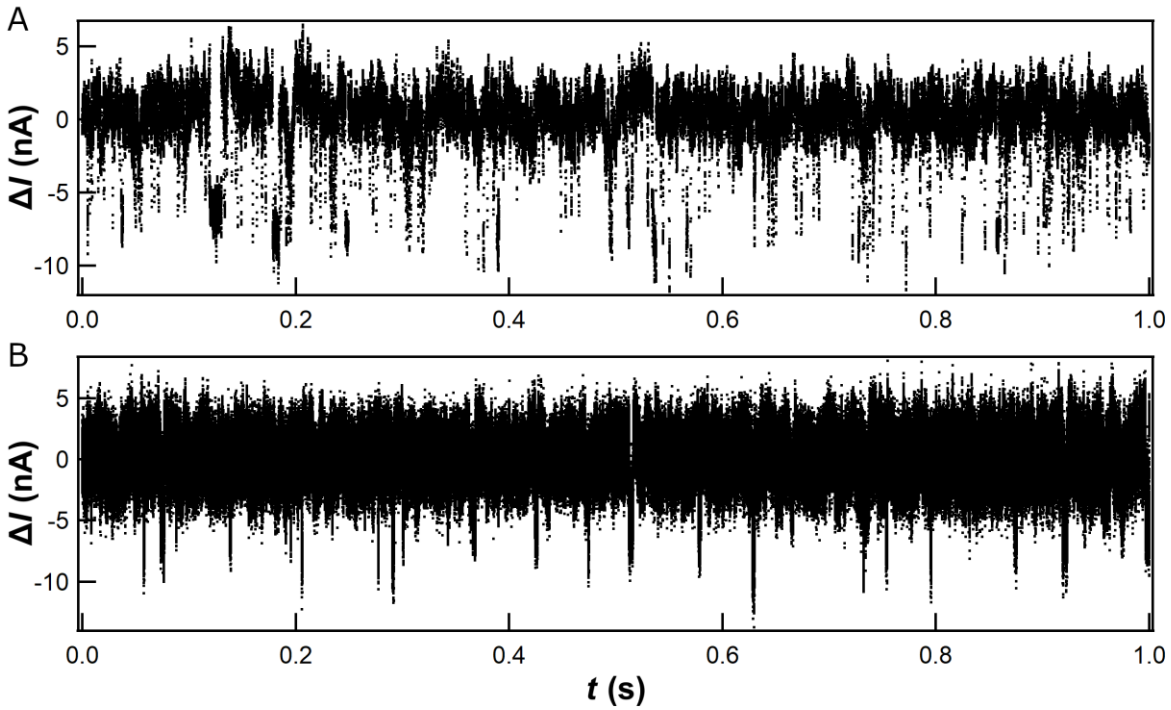

Figure 4.1: Examples of the $\Delta I(t)$ acquired at (A) 25 kHz and (B) 220 kHz bandwidth.

Digital noise reduction, the process of reducing or removing unwanted noise from an acquired electronic signal, is an important segment of signal processing, with applications in a variety of fields such as audio (137) and visual (138, 139) communications, chemical (140) and medical sensing (141, 142), and seismology (143, 144). There are multiple approaches for noise reduction in a digital signal. The most common schemes utilize some type of digital filter to selectively attenuate noise while preserving most of the desired signal. Some schemes operate conceptually in the time or space domain (such as the median filter (145), total variation denoising (146), or Savitzky–Golay filter (147, 148)), while others operate in the frequency domain (infinite impulse response (IIR) filters), and some operate in both domains simultaneously (wavelet denoising (149)).

Effective denoising occurs when the operating domain of the denoising scheme matches the domain of the important features in the signal. For instance, audio denoising works best in the frequency domain because the relevant features tend to be waves (or at least wavelets), suggesting the use of IIR or FIR (finite impulse response) filters to separate signal frequencies from the remaining noise (137, 150). However, visual image denoising prioritizes image smoothness and the location of discrete edges, both of which are features of the space domain (151). Thus, median filtering or total variation denoising are better suited for denoising images (145, 146). Since frequency-based schemes are inherently dependent on sinusoids, such schemes tend to smooth sharp edges and blur images.

In the SWCNT-FET biosensor, the signals that correspond to enzyme conformational changes are often short-duration spikes on a mostly low-frequency noise background (1-3, 25). This presents a denoising problem that is significantly different from fields such as audio or visual image denoising, which are often concerned with removing high-frequency portions of white noise or discrete salt-and-pepper noise (which are effectively sharp spikes) on an otherwise lower-frequency signal. Most applications of denoising procedures intend to remove sharp spikes, but with enzyme conformational changes the sharp spikes are in fact the signal of interest. This suggests that a different approach is needed to extract the signal corresponding to enzyme activity from the noise in a SWCNT-FET biosensor.

In the past few decades, new signal processing concepts and advances in computation power facilitated new approaches for separating a desired signal from noise. A major advance was the development of wavelet theory and the wavelet transform, which performs analysis of a

signal in both time and frequency domains. Noise reduction using the wavelet transform is used in image (138, 151-153) and audio (137, 154) denoising, artifact removal in EEG (142) and ECG (141, 155, 156) recordings, chemical detection (157-159), and seismology (144).

The following is a brief description of wavelet denoising as applied to SWCNT-FET signals – a detailed explanation of the wavelet transform is beyond the scope of this work. Additional information on the wavelet transform can be found in the references (149, 160-162).

4.2 Multiresolution Analysis and the Undecimated Wavelet Transform

4.2.1 Wavelet Transform Fundamentals

The traditional way to analyze a time series signal in the frequency domain is to calculate the Fourier transform of the signal, then look for notable features in the resulting spectrum (149, 163). The basis function of the Fourier transform, a sinusoid, is well-localized in the frequency domain but completely delocalized in the time domain. This means the Fourier transform is an excellent tool for characterizing signals which are periodic or whose frequency content varies slowly with time, but it is inadequate for properly characterizing signals which exhibit stochastic behavior in the time domain. In addition, any sharp spikes or transitions that are well-localized in the time domain will appear in multiple frequency levels in the frequency domain, due to the inherent time-frequency uncertainty relation. Although modifications such as the short-time Fourier transform attempt to address these issues (164, 165), these modifications keep the fixed time and frequency resolutions of the

original Fourier transform, resulting in worse-than-optimal resolution at low frequencies and short times.

The wavelet transform is a time-frequency transform, analogous to the Fourier transform, which uses a wavelet instead of a sinusoidal function as its basis (149, 162). Wavelets are wave-like functions or patterns localized in both the time and frequency domains (within the bounds set by the uncertainty principle). In contrast to the Fourier transform, which only produces a 1D spectrum of frequency amplitudes, the localization of the wavelet transform in both domains allows the wavelet transform to meaningfully describe a signal both in time and frequency, producing a 2D plot of coefficients. Figure 4.2 shows how the wavelet transform divides the time-frequency space into blocks whose dimensions maximize information content: at low frequencies, blocks are long in time to facilitate high frequency resolution, while at high frequencies, blocks are short in time to facilitate high time resolution. The area of each block is identical, with a minimum defined by the wavelet basis and the time-frequency uncertainty principle (149). The primary advantage of the wavelet analysis is the ability to localize changes in a signal in both time and frequency, which includes any sharp transitions or spikes in the signal.

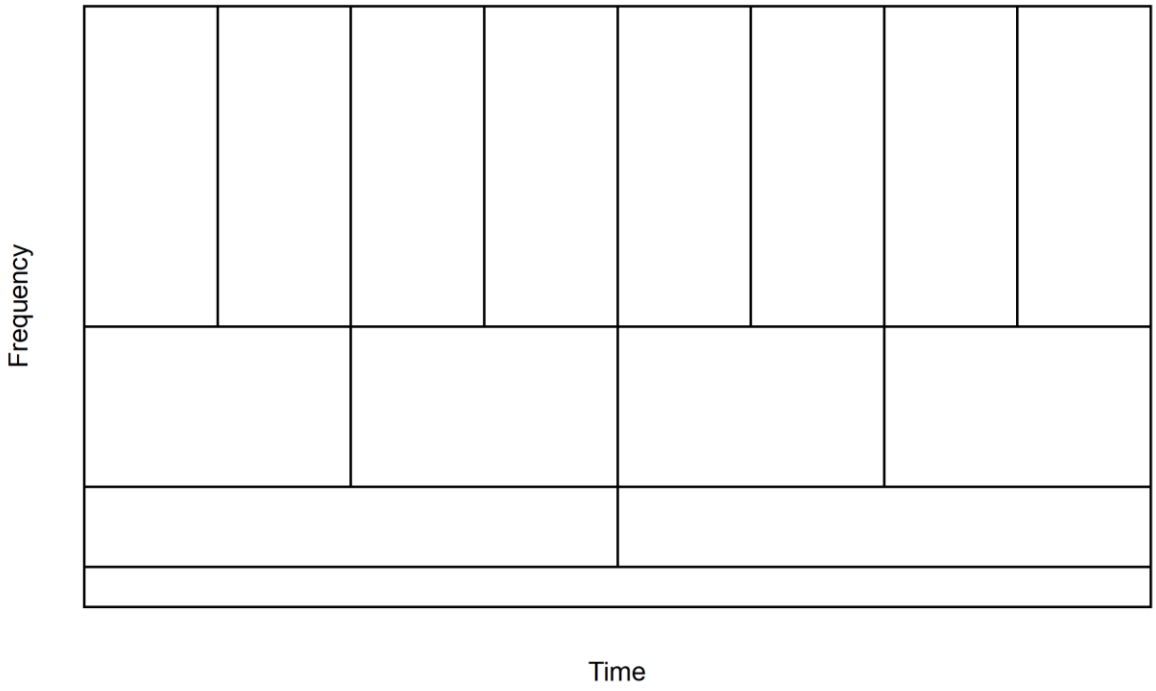

Figure 4.2: Division of the time-frequency space in a wavelet transform representation. Both the time and frequency axes are displayed on a linear scale.

Conceptually, the 2D plot of wavelet coefficients is constructed by convolving the wavelet function $\psi_{s,b}(t)$ with the signal $x(t)$ over all values of scale s, or dilations of the original wavelet function. The scale, which is analogous to inverse frequency $\left(s \propto \frac{1}{f}\right)$, describes the amount of stretch (in time) applied to the wavelet before the convolution. The general wavelet transform is illustrated in Figure 4.3. $\psi_{s,b}(t)$ (red, left) is scanned over $x(t)$ to perform the convolution, then the wavelet function is stretched by s (blue, right), and the convolution repeated. Mathematically, this is expressed as:

$$\psi_{s,b}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-b}{s}\right)$$

$$X_w(s,b) = \int_{-\infty}^{\infty} x(t)\,\psi_{s,b}\,dt = \frac{1}{\sqrt{s}}\int_{-\infty}^{\infty} x(t)\,\psi\left(\frac{t-b}{s}\right)dt$$

where $X_w(s, b)$ is the wavelet coefficient for a particular s and translation factor b. This expression exactly describes the wavelet transform when applied to continuous functions or signals (and such a transform is termed the continuous wavelet transform, or CWT).

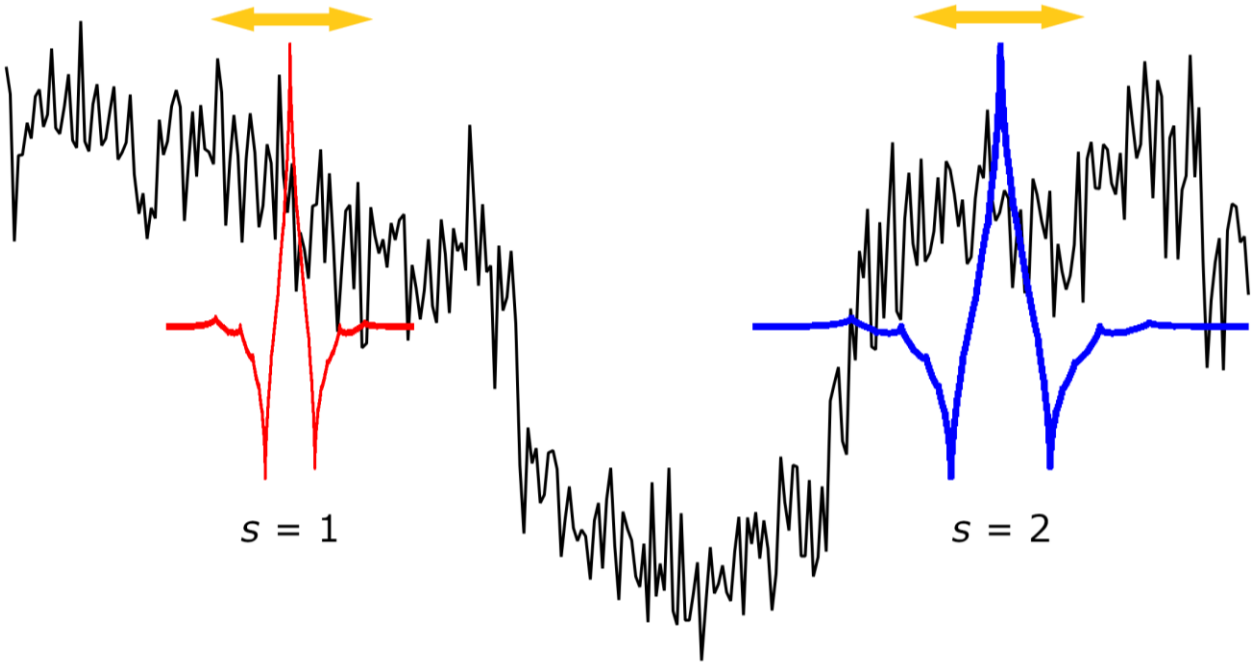

Figure 4.3: Calculating the wavelet transform of a 1D signal. The wavelet function (s = 1) is convoluted with the signal starting at t=0, then stretched in time (s = 2) and convoluted again with the signal, to produce a 2D array of coefficients.

4.2.2 The Discrete and Undecimated Wavelet Transforms

However, for discrete signals, the integrals are converted to sums, and both s and b are restricted to integer values. The resulting transform is called the discrete wavelet transform, or DWT. Each value of s corresponds to an independent frequency band, with the highest frequencies corresponding to the lowest scales and the lowest frequencies corresponding to the highest scales.

Although s can take any integer value, essentially all implementations of the DWT use a dyadic scaling, selecting only powers of 2 such that

$$s = 2^n, 0 \leq n \leq \log_2 N\,, n \in \mathbb{Z}, n > 0$$

where N is the number of samples in $x(t)$ and n is an integer greater than 0. Dyadic scaling presents several advantages: 1) the resulting wavelet functions are orthonormal, such that the information content in each scale is independent of the others, 2) the transform provides logarithmic coverage of the frequency space, requiring less coefficients to describe the entire space, and 3) scaling by powers of 2 is computationally inexpensive.

In fact, the dyadic scaling of the wavelet transform means that the transform coefficients can be calculated more efficiently. Due to the dyadic stretching of $\psi_{s,b}(t)$, each scale in the CWT contains half the frequency information of the previous scale but the same number of coefficients, resulting in a redundant representation of the information. Halving the number of coefficients at each successive scale reduces memory and computing requirements while preserving the frequency information. An alternative approach, instead of calculating the convolution equally at every scale, passes $x(t)$ through a pair of quadrature mirror filters (acting as high- and low-pass filters), which are designed such that the filters reproduce the output of the DWT for a given $\psi_{s,b}(t)$. The output of the high-pass filter, named the detail coefficients, are saved as the entire first scale level of the wavelet transform, while the output of the low-pass filter, named the approximation coefficients, is decimated by 2. When the decimated approximation coefficients are passed through the high- and low-pass filters again, the resulting detail coefficients are saved as the entire second scale level (now with

half as many coefficients). A general schematic of this process is shown in Figure 4.4. This algorithm is repeated until the approximation coefficients cannot be decimated further, resulting in a series of scale levels in which each level contains half as many coefficients as the previous level.

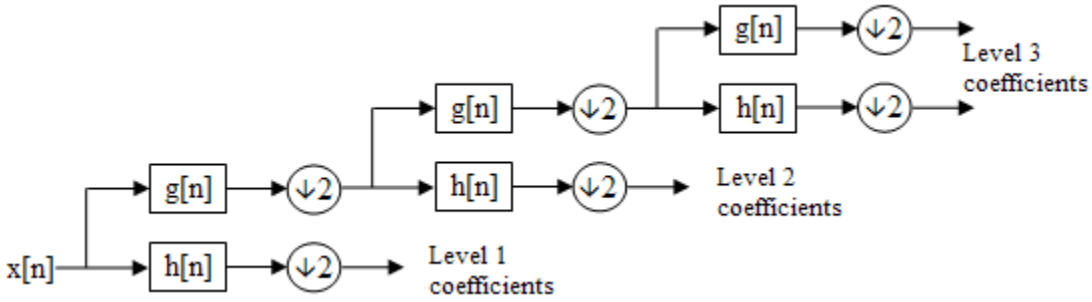

Figure 4.4: Schematic of the discrete wavelet transform, showing the high/low pass filtering and the decimation by 2 at each step. The undecimated wavelet transform does not perform the decimation.

The primary disadvantage of the dyadic DWT algorithm is that it is not translation-invariant – shifting $x(t)$ in time changes the amplitudes of the coefficients in addition to shifting their position in time. The lack of translation invariance can be a significant problem if the purpose of the wavelet transform is to identify the specific locations of interesting features in $x(t)$, such as the time of a sharp transition or spike. A solution is to use a translation-invariant version of the DWT, such as the undecimated wavelet transform (UWT) (166, 167), which follows the same algorithm as the DWT but omits the decimation by 2. The series of scale levels in the UWT contains more coefficients (and requires more computation) than in the DWT because the levels are not decimated, but this allows the UWT coefficients to maintain both translation invariance and location accuracy of interesting signal features.

4.2.3 Selecting a Wavelet Basis

The selection of wavelet basis is an important consideration in optimizing the use of the wavelet transform. Conceptually, the wavelet transform performs best in identifying and localizing features that are similar in shape to the wavelet itself. Though almost any oscillating, zero-mean, finite-energy wavelet could potentially be used in the wavelet transform, there are certain families of wavelets that have been developed with useful properties (149, 168, 169). A wavelet with many oscillations (like the Morlet for the CWT or the higher-order Daubechies for the DWT) is most useful for identifying periodic oscillations, while a wavelet with sharp edges (like the Haar for the DWT) is best suited for finding sharp transitions in time, and a wavelet with sharp spikes (such as the Coiflet family for the DWT) excels at finding the location of sharp spikes in $x(t)$ (170, 171). Some examples of wavelet bases are displayed in Figure 4.5. Most designed wavelet families are orthogonal, so frequency information is not shared among multiple scales, and some wavelet families are symmetric.

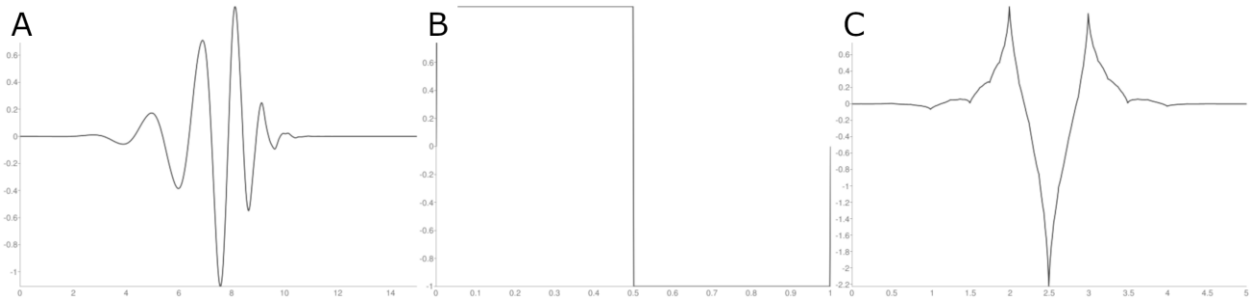

Figure 4.5: Examples of wavelet bases: (A) Daubechies 8, (B) Haar, and (C) Coiflet 1

4.3 Wavelet Thresholding and Denoising

4.3.1 Decorrelating 1/f Noise

The predominant source of electrical noise in carbon-based electronics is flicker noise, also called 1/f noise because the power spectral density of flicker noise is approximately inversely proportional to frequency:

$$PSD \propto \frac{1}{f^{\beta}}$$

where $1.0 \leq \beta \leq 1.1$ for carbon nanotubes specifically (29). 1/f noise is ubiquitous, appearing in fields as diverse as electronic noise, economic data, and biological processes (172-174). Despite its ubiquity, removing 1/f noise from signals is a challenging problem. One of the primary issues in dealing with this type of noise is its inherent long-range correlation, which prevents the signal from converging or averaging to some fixed mean value. Instead, a signal with 1/f noise has an unpredictable, long-term drift in mean value that confuses level-finding algorithms. Most approaches for denoising involve high-pass filtering to reduce the amplitude of low-frequencies, where the 1/f noise has the most power, or else using some sort of chopping or modulation to push the noise outside the bandwidth of any desired signals (175). Though this approach is effective for many applications, high-pass filtering removes potentially important low-frequency information and generates artifacts around sharp transitions. A better approach involves selectively removing 1/f noise at all frequencies.

One property of 1/f processes that can be exploited to facilitate effective denoising at all frequencies is its statistical self-similarity, meaning that the characteristics of the process appear the same when measured on different timescales (160, 161). This means that such processes do not possess any "characteristic" timescales or correlation lengths, and thus cannot be accurately described by traditional correlation or time-based metrics. Instead, proper analyses of 1/f processes should use methods that are scale- and translation-invariant. As already shown above, the discrete wavelet transform meets both criteria, making it a perfect tool to characterize 1/f noise. In particular, when 1/f noise is passed through the DWT, the coefficients at an individual scale level become mutually uncorrelated, effectively "whitening" the noise and facilitating easier removal through traditional white noise reduction methods. This whitening of 1/f noise is made possible because of two factors: 1) the orthonormality of the basis functions used in the discrete wavelet transform prevents noise information passing between scale levels, and 2) the bandpass filters act on a logarithmic frequency axis, neatly separating the frequency space into self-similar bands.

Certain statistical metrics of 1/f noise are invariant across scale levels. One such metric is the variance of the coefficients within a single scale level, with the invariance described below for any scale factor *s*:

$$Var\ X^s = \sigma^2 2^s$$

$$Std.Dev.X^s = \sqrt{Var\ X^s} = 2^{\frac{s}{2}}\sigma = \left(2^{\frac{1}{2}}\right)^s \sigma$$

where $\frac{\sigma}{\sqrt{2}}$ is the standard deviation of the first scale level.

From the PSD of 1/f noise, the amplitude of the noise at a particular frequency scales as:

$$PSD \propto \frac{1}{f} \propto Amp^2$$

$$Amp \propto \frac{1}{f^{\frac{1}{2}}}$$

The ratio of frequencies between two adjacent scale levels in the DWT is 2, so the ratio of the amplitudes of the noise between adjacent scale levels is:

$$\frac{Amp_{s+1}}{Amp_s} = \frac{\frac{1}{f_{s+1}^{\frac{1}{2}}}}{\frac{1}{f_s^{\frac{1}{2}}}} = \frac{\frac{1}{f^{\frac{1}{2}}}}{\frac{1}{(2f)^{\frac{1}{2}}}} = \frac{(2f)^{\frac{1}{2}}}{f^{\frac{1}{2}}} = 2^{\frac{1}{2}}$$

which is the exact same factor expressed in the standard deviation of each scale level. Thus, by reducing the amplitude of the coefficients in each scale level by $\left(2^{\frac{1}{2}}\right)^s$, the noise amplitude can be reduced to the same level across all frequencies, essentially turning 1/f noise into uncorrelated white noise. The resulting coefficients can be further denoised using methods suitable for white noise.

4.3.2 Wavelet Thresholding

Wavelet thresholding, sometimes called wavelet shrinkage denoising, is a non-linear denoising technique performed on the wavelet coefficients (138, 176, 177). The basic assumption in wavelet thresholding is that wavelet coefficients associated with noise have smaller amplitudes than the coefficients associated with the signal of interest. Thus, by removing all coefficients below a certain threshold (replacing the coefficient values with 0),

the noise can be removed while preserving the desired signal. Wavelet denoising works best when the wavelet basis matches the important features of the desired signal, such that the feature can be accurately described with only a few wavelet coefficients (170, 176). In the ideal case, this produces a sparse representation in the wavelet domain, in which the significant features have non-zero coefficients and the remaining coefficients are zero.

The choice of threshold is crucial for optimizing the signal-to-noise ratio of the resulting denoised signal. There are a variety of schemes for choosing the threshold. Some apply a universal threshold across all scale levels, while others vary the threshold according to the scale level. In addition, there are several methods for handling the coefficients that are larger than the threshold (153). A "hard" threshold leaves such coefficients unchanged, resulting in a signal that preserves sharp edges but may also be more uneven and jagged. A "soft" threshold (178) shrinks all coefficients by the threshold value, resulting in a smoother signal. A "garrote" threshold (179) is an intermediate between the "hard" and "soft" threshold methods, shrinking coefficient values that are just above the threshold while leaving large coefficient values (greater than 5 times the threshold value) mostly unchanged. A schematic of the different types of thresholding, with the threshold value set to 1 in each case, is shown in Figure 4.6. An example of both soft (blue) and hard (red) thresholding applied to a sample scale level is shown in Figure 4.7. In this example, both methods utilize the same threshold value. When the original coefficient is higher than the threshold, the hard threshold method does not alter its value and preserves the peak amplitude, but the soft threshold shrinks the coefficient value by the threshold value and reduces the amplitude of the resulting peak.

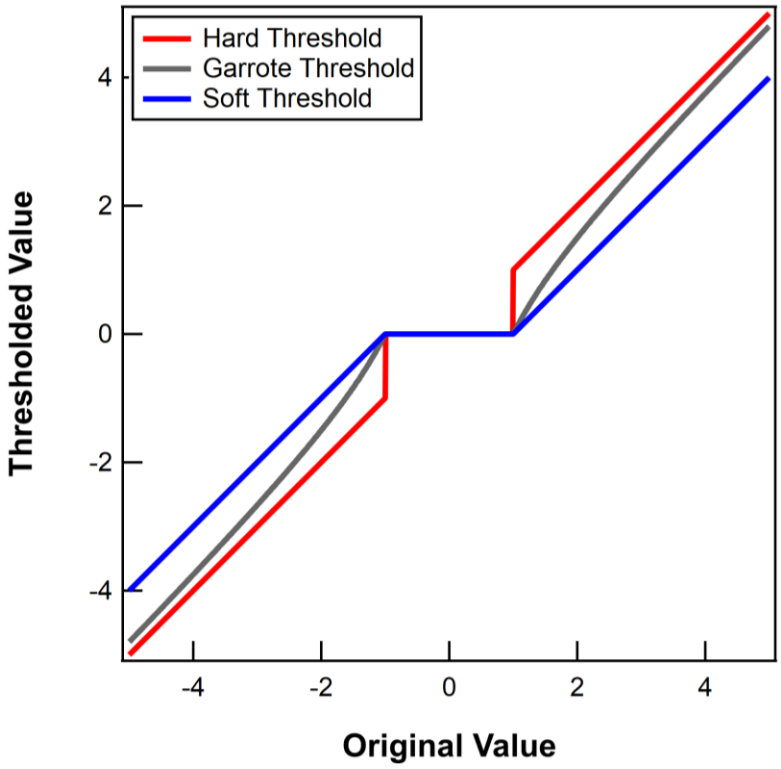

Figure 4.6: The relationship between the original and thresholded values for hard, garrote, and soft thresholding. Here, the threshold value is 1.

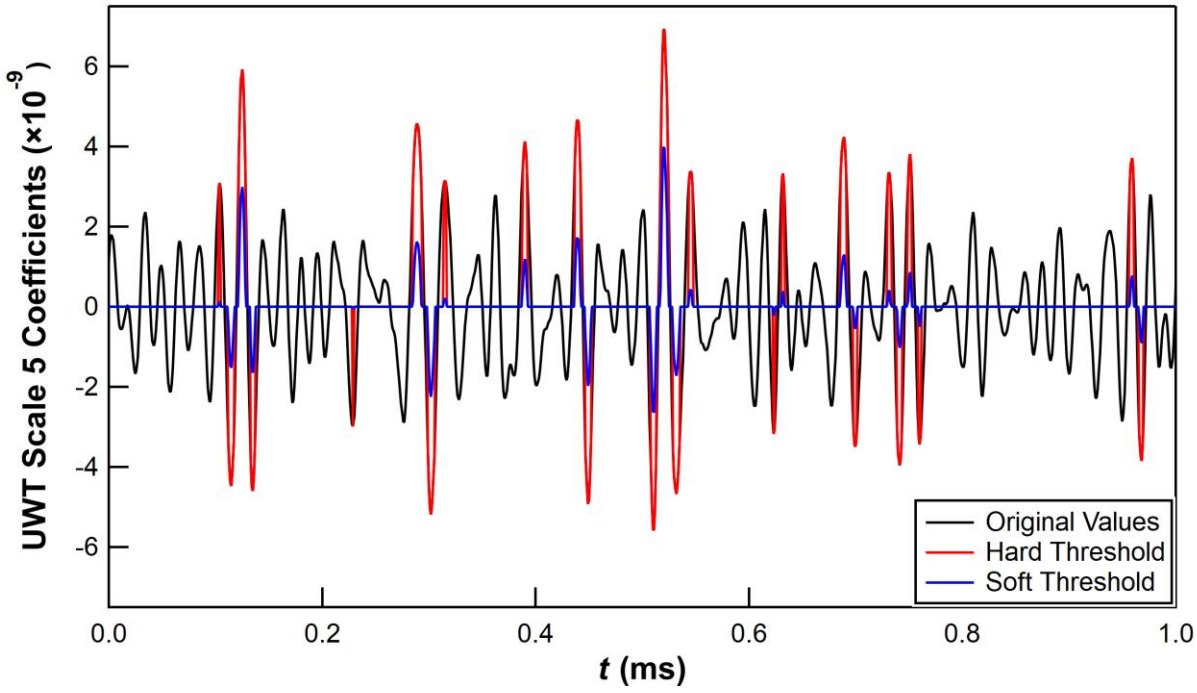

Figure 4.7: Results of hard (red) and soft (blue) thresholding of the coefficients from a single scale in the UWT, with the original coefficients in black.

The choice of thresholding scheme is heavily dependent on the type of application and on the characteristics of the main signal and the noise. For instance, many soft thresholding schemes have been proposed by researchers for visual image denoising (138, 153, 180, 181), where smoothness is generally prioritized over edge sharpness. Hard thresholding is not widely utilized, but tends to perform better when edge sharpness is the highest priority or when the magnitude of the original signal needs to be preserved.

4.4 Wavelet Denoising of the SWCNT-FET Electrical Signal

4.4.1 Scale-Dependent Thresholding of UWT Coefficients

In the SWCNT-FET biosensor, 1/f noise from the carbon nanotube is mixed together with other types of electrical noise from various sources. These include: high-frequency noise from switching power supplies, broadband noise from the liquid heating control circuitry, and broadband noise from building electronics, fluctuations in grounding, and even other experiments in the building. These additional noise sources add to the total noise in the SWCNT circuit and appear either at specific frequencies (resulting in spikes in the plot of the power spectral density (PSD), as shown on the right of Figure 4.8) or as a broad band that increases the noise floor of the measurement. At a certain frequency (~200 kHz in Figure 4.8), the amplitude of the noise floor becomes larger than the 1/f noise from the nanotube itself. This point, called the corner frequency, creates a natural divide in the frequency space.

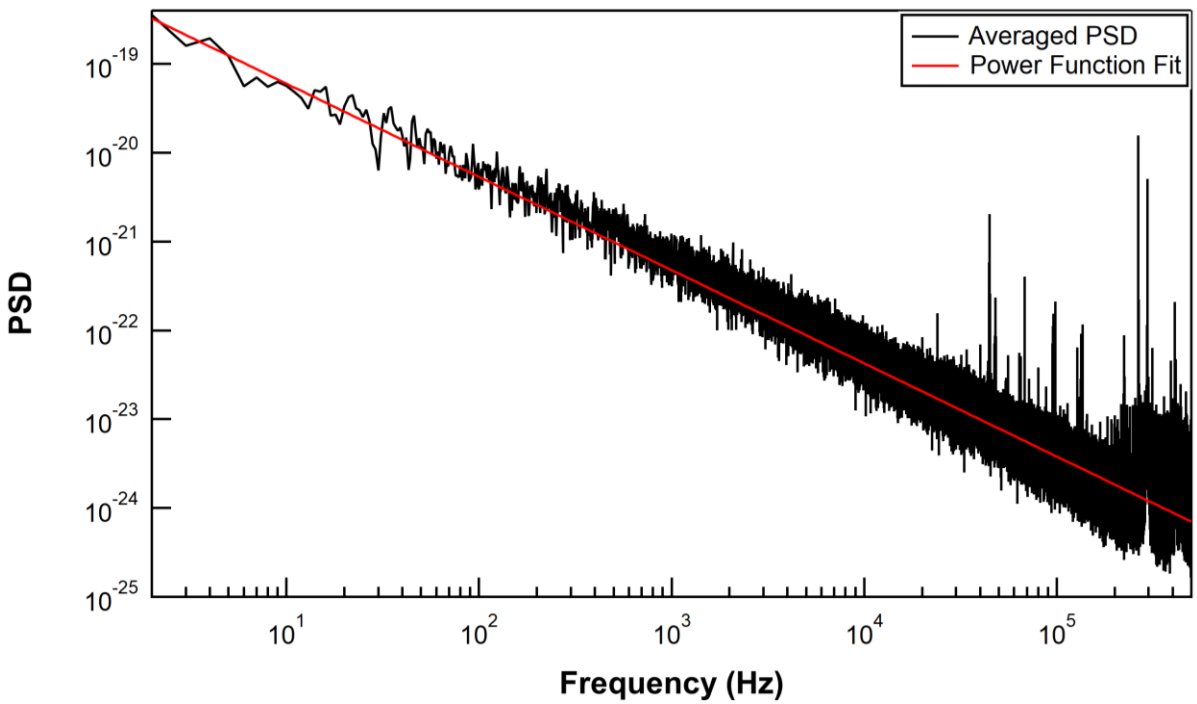

Figure 4.8: PSD of a 1s segment of signal from the SWCNT-FET. Note that the slope of the data is ~-1 on a log-log plot, corresponding to a function proportional to 1/f.

Since the SWCNT-FET signal contains a variety of frequency-dependent noise sources, the optimal thresholding approach uses a different threshold for each scale. The noise amplitude at each scale is calculated using the median absolute deviation (MAD), which is defined by:

$$MAD(x(t)) = median\big(|x(t) - median(x(t))|\big)$$

The MAD is an analogue to the standard deviation that is not as sensitive to outliers. This is useful since significant spikes in the signal are correlated with large coefficients, which can be considered "outliers" to the 1/f noise. Thus, the MAD provides a metric for the noise amplitude that is not influenced by the amplitude of the spikes in the signal. Figure 4.9 displays a plot of the MAD of the coefficients at each scale level, along with a line showing the exponential relation: $MAD \propto \left(2^{\frac{1}{2}}\right)^{s}$, which is the same characteristic relation mentioned above for 1/f noise. The plot reveals a close agreement between the characteristic relation

and the experimentally-obtained MAD at each scale, with a slight difference in slope because, for SWCNT-FETs, $\beta > 1$ in the characteristic equation defining 1/f noise: $PSD \propto \frac{1}{f^{\beta}}$.

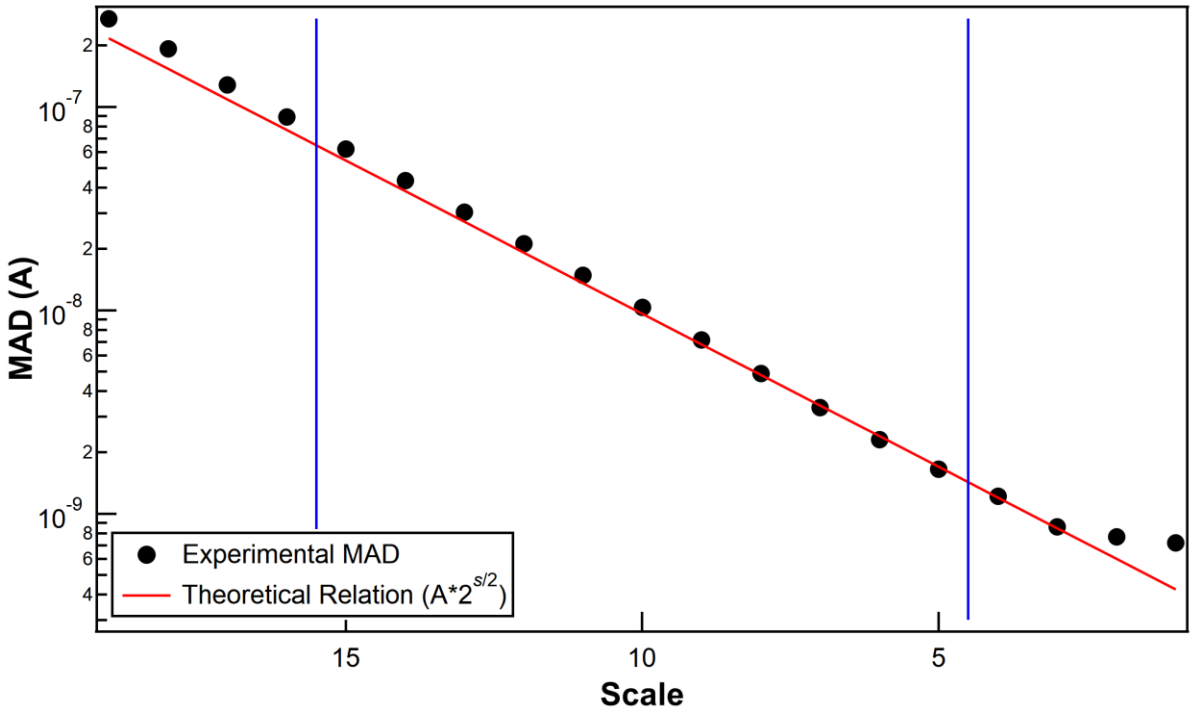

Figure 4.9: Comparison of the MAD of the coefficients at each scale (black dots) to an exponential function (red) the amplitude scaling of pure 1/f noise. Blue lines indicate the divide between the low- (left), mid- (center), and high- (right) frequency regions.

With 10^6 samples in a single batch (corresponding to 1 s of 1 MHz signal), the wavelet transform representation contains 19 scale levels, since $\log_2 N = \log_2 10^6 \approx 19.93$. The frequency space is divided into three sections: a high-frequency region (scales 1-4), a mid-frequency region (scales 5-15), and a low-frequency region (scales 16-19). The separations between these sections are shown in Figure 4.9 as blue vertical lines. Within each section, the UWT coefficients are thresholded in a manner that depends on scale and three additional parameters: threshold type (hard, soft, garrote), threshold level, and number of scales.

In the following threshold calculations, the MAD is used as an analogue to the standard deviation. The distribution of the UWT coefficients within a single scale can be approximated by a normal distribution, which has a well-defined percentage of outliers beyond a certain number of standard deviations. For example, using a threshold of 2 standard deviations preserves only values above the 95th percentile, while a threshold of 4 standard deviations corresponds to keeping only values above the 99.99th percentile. Similar thresholding can be achieved by replacing the number of standard deviations with a multiplier of the MAD. Thus, a larger multiplier removes greater percentages of 1/f noise but may also suppress signal dynamics if the corresponding coefficient amplitudes are smaller than the threshold.

For the scales in the "mid-frequency" region, below the corner frequency and dominated by 1/f noise, the coefficients are most optimally denoised using a scale-dependent threshold. Since the standard deviation of the noise scales as $\left(2^{\frac{1}{2}}\right)^{s}$, a threshold that also scales in the same manner will remove similar proportions of noise at each scale level. The threshold for a scale *s* in this region is calculated from:

$$t_{s,MF} = 2^{\frac{s_m - s}{2}} * (MAD)_m * m_{MF}$$

where s_m is the number of the middle scale (for 19 scales, scale 10 is the middle), $(MAD)_m$ is the MAD of the coefficients in the middle scale, and m_{MF} is the multiplier for the threshold (usually set between 2-4). Hard thresholding is used to preserve sharp transitions.

In the low-frequency region, the main objective is to remove fluctuations that are too slow for biomolecule activity, making the baseline flat and keep a consistent position from one batch of signal to the next. The approximation coefficients are all set to 0 to make the denoised signal have a mean of zero. Since each batch of signal is 1 s long, zeroing these coefficients essentially removes frequencies below 1 Hz. In addition, the details in the lowest-frequency scales contain essentially no information related to biomolecule activity, so the coefficients in these scales can be set to 0 as well, which essentially sets the threshold to +∞. The largest n_{LF}, the number of scales included in the low-frequency section, is 4, which corresponds to scales 16-19 and characteristic times τ > 66 ms when operating on 1 MHz data. This results in a high-pass filter with a ~15 Hz cutoff. Using $n_{LF} > 4$ distorts the signals from long-duration biomolecule activity, which can be up to ~10 ms in duration.

In the high-frequency region, beyond the corner frequency, fast biomolecule dynamics are often masked by instrumentation noise, so only transitions or spikes with amplitudes larger than the noise can be identified. In this frequency range, hard thresholding is used if the denoising needs to preserve signal timescales that are within two orders of magnitude of the sampling interval (for example: spike durations < 100 μs when the sampling interval is 1 μs for 1 MHz data). Otherwise, garrote thresholding is used. The threshold is determined by:

$$t_{s,HF} = (MAD)_s * m_{HF}$$

where $(MAD)_s$ is the MAD of the coefficients in scale level *s*, and m_{HF} is the multiplier for the threshold, set between 2 and 6. The number of scales n_{HF} included in this section depends on both the corner frequency and minimum signal timescale, and ranges from 4 (scales 1-4, corresponding to τ < 16 μs) to 8 (scales 1-8, corresponding to τ < 256 μs) for 1 MHz data.

4.4.2 Multi-Channel Denoising For Multiple Timescales

Generally, signals corresponding to biomolecule activity are present over a wide range of timescales that may span several orders of magnitude. This is due to several factors. First, biomolecules may possess multiple functions or conformational changes, each of which may operate at a different timescale (35, 106). Second, biomolecule motion generally follows Poisson statistics, so the amount of time a biomolecule stays in a single conformation has an exponential probability distribution over several orders of magnitude in time (39). For example, a particular biomolecule conformation might have a characteristic time of τ = 300 μs, but the biomolecule will exhibit conformation dwell times ranging from 60-600 μs. Finally, biomolecule activity is also subject to dynamic disorder (40, 182), which are changes in the conformational energy landscape of the biomolecule that alter the characteristic times of conformational states over time, sometimes by an order of magnitude.

The wide range of possible timescales present in the SWCNT-FET signal presents a problem for wavelet 1/f denoising, since the scale-dependent thresholds are effective only within a range of timescales, between τ_{min} and τ_{max}. For timescales shorter than τ_{min}, sharp spikes are smoothed away and sharp transitions rounded, while for timescales longer than τ_{max}, rectangular steps are flattened. If the timescale range of biomolecule activity is larger than the timescale range of a single denoising channel, then multiple denoising channels are required, each spanning a different range of timescales, so that all the timescales are covered. Using multiple denoising channels also helps isolate signal components arising from different conformational motions, especially if those motions exist in separate timescales.

For the purposes of denoising SWCNT-FET signals arising from measurements of Taq DNA polymerase, two channels are used: a "low-pass" channel and a "high-pass" channel. Note that these are not true low-pass and high-pass filters – both channels are bandpass filters. The "high-pass" channel preserves shorter timescales than the "low-pass" channel, while the latter channel produces a smoother output which makes it easier to find edges in the signal. A list of the channels for denoising 1 MHz recordings of Taq DNA polymerase, along with the denoising parameters and targeted timescales, is shown in Table 4.1.

Denoising Channel	Low Freq.	Mid Freq.		High Freq.			τ_{min} (μs)	τ_{max} (ms)
	Included Scales	Thresh. Type	m_{MF}	Included Scales	Thresh. Type	m_{HF}		
High-Pass	16-19	Hard	3	1-8	Hard	2	1	.5
Low-Pass	16-19	Hard	4	1-4	Garrote	6	50	10

Table 4.1: Channels and parameters for wavelet denoising of the SWCNT-FET signal.

Wavelet denoising was implemented in LabVIEW using the following sequence of steps:

1) separate the signal into batches (usually 1s of data, or 10^6 samples at 1MHz),
2) decompose the current batch of signal into its undecimated wavelet transform (UWT) representation,
3) for each denoising channel, threshold the UWT coefficients according to the thresholding method and value at each scale,
4) for each denoising channel, generate the denoised $I(t)$ signal by applying the inverse undecimated wavelet transform (iUWT) on the thresholded scales.

A schematic for the entire wavelet denoising scheme, including the multi-channel approach, is shown in Figure 4.10.

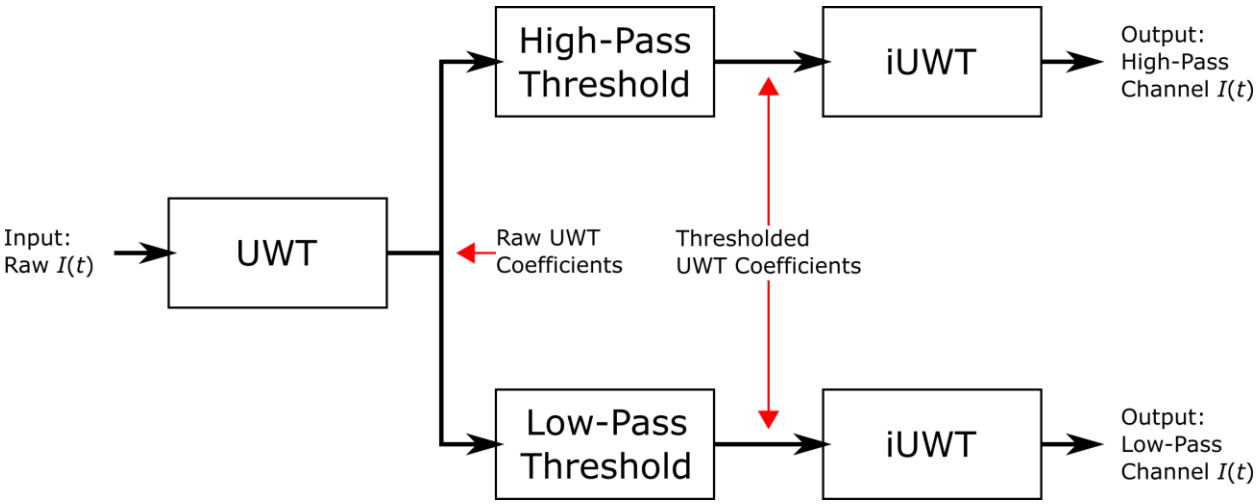

Figure 4.10: Block diagram of the multi-channel wavelet denoising scheme.

To illustrate the performance of this wavelet denoising procedure over a range of timescales, Figure 4.11 compares the raw $I(t)$ (black), obtained from SWCNT-FET recordings of Taq DNA polymerase activity, to the denoised output of the low-pass channel (red), at timescales ranging from 60 s (top) to 2 ms (bottom). In Figure 4.11A, the 60 s segment of raw signal exhibits low-frequency fluctuations of ±5 nA in the baseline, while in the denoised signal the baseline appears flat and fluctuates by less than 2 nA. The standard deviation of the signal drops from 2.9 nA for the raw signal to 0.7 nA for the low-pass channel. At this scale, the spikes corresponding to individual conformation changes are impossible to distinguish. Snapshots of the signal at progressively smaller timescales show that the denoised signal consists of individual spikes or clusters of spikes separated by a smooth, often flat, baseline.

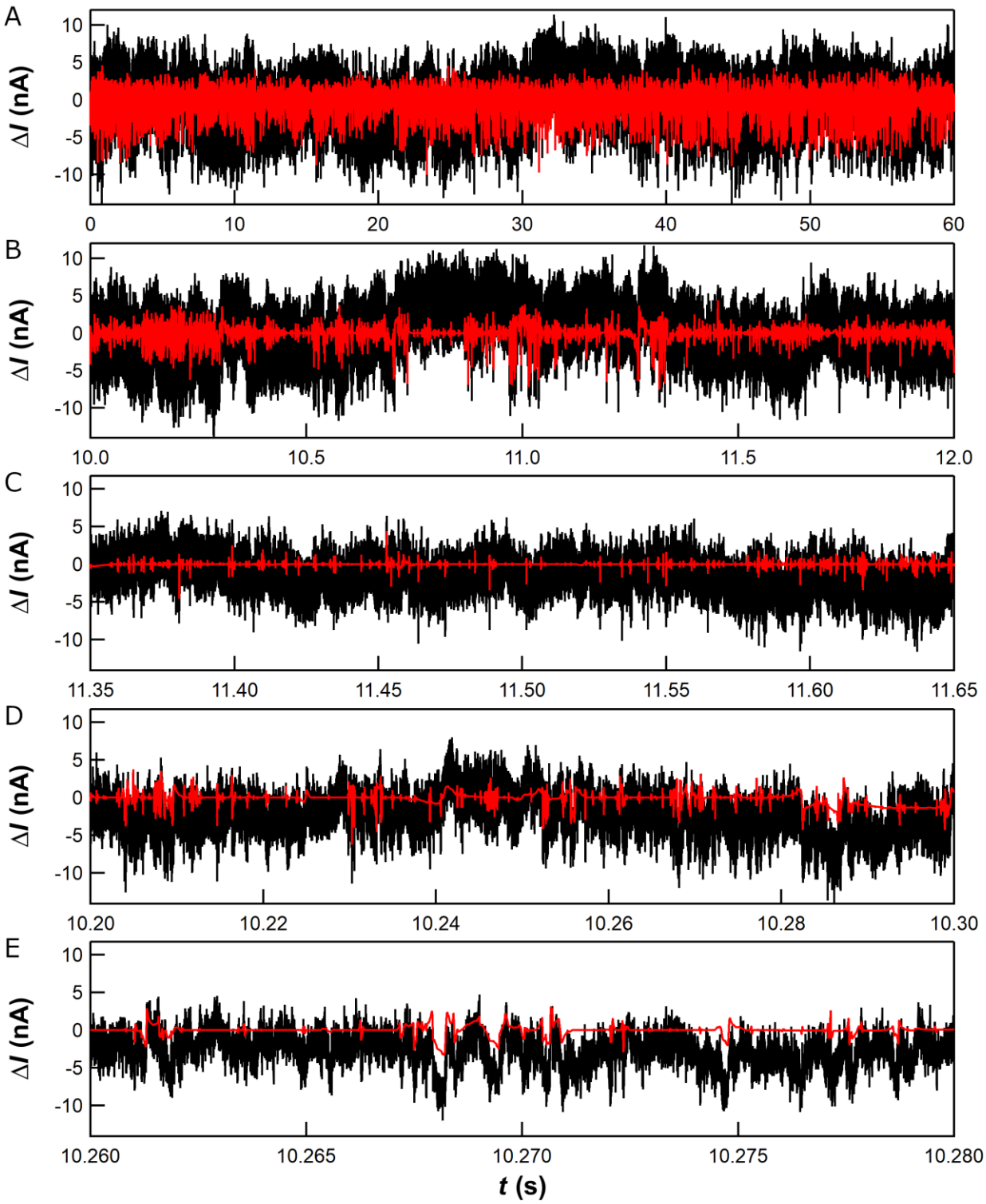

Figure 4.11: Examples of raw (black) and low-pass wavelet-denoised (red) signal from the SWCNT-FET, at a variety of timescales: (A) 60 s, (B) 2 s, (C) 300 ms, (D) 100 ms, and (E) 20 ms.

Sometimes, the low-pass denoised signal is flat even when the raw signal seems to have large fluctuations, such as at t = 11.5 s in Figure 4.11B. Figure 4.11C zooms in on a 300 ms snapshot of the signal at that time, showing that most of the fluctuations are gradual transitions due to 1/f noise rather than sharp transitions due to the biomolecule. There are some sharp spikes present in this segment of signal, but the amplitudes of the spikes are small relative to the 1/f noise peak-to-peak (signal to noise ratio (SNR) ~1.5:1), so the spike amplitudes are reduced in the denoised signal due to garrote thresholding. At other times, the denoised signal seems to exhibit dense clusters of spikes, like at t = 10.2 s in Figure 4.11B. Figure 4.11D zooms in on 100 ms of signal at that time, showing that there are still gaps of flat baseline signal between most spikes, even in a cluster. Some of the spikes have their amplitudes reduced by more than 50% in the denoised signal, and some spikes are smoothed away entirely. Zooming in further to 20 ms (Figure 4.11E) shows that the spikes that are completely smoothed away either have a low SNR (< 2) or have gradual rather than sharp transitions (such as the triangular-looking wave at t = 10.268 s).

Figure 4.12A zooms in on another 100 ms snapshot of the signal (taken from the same segment displayed in Figure 4.11B) where the low-pass denoised signal preserves > 70% of the spike amplitudes. Figure 4.12B zooms in further to a 10 ms snapshot, showing that these spikes exhibit higher SNR (~3) and have sharper transitions. At another 15 ms snapshot (Figure 4.12C), the amplitudes of some spikes or transitions are completely preserved (like at t = 11.0325 s and 11.038 s), while the amplitude of others are reduced (like at t = 10.0315s). Zooming in further on these times (Figure 4.12D and E) show that spikes with durations < 100 μs are rounded off or smoothed away to some degree.

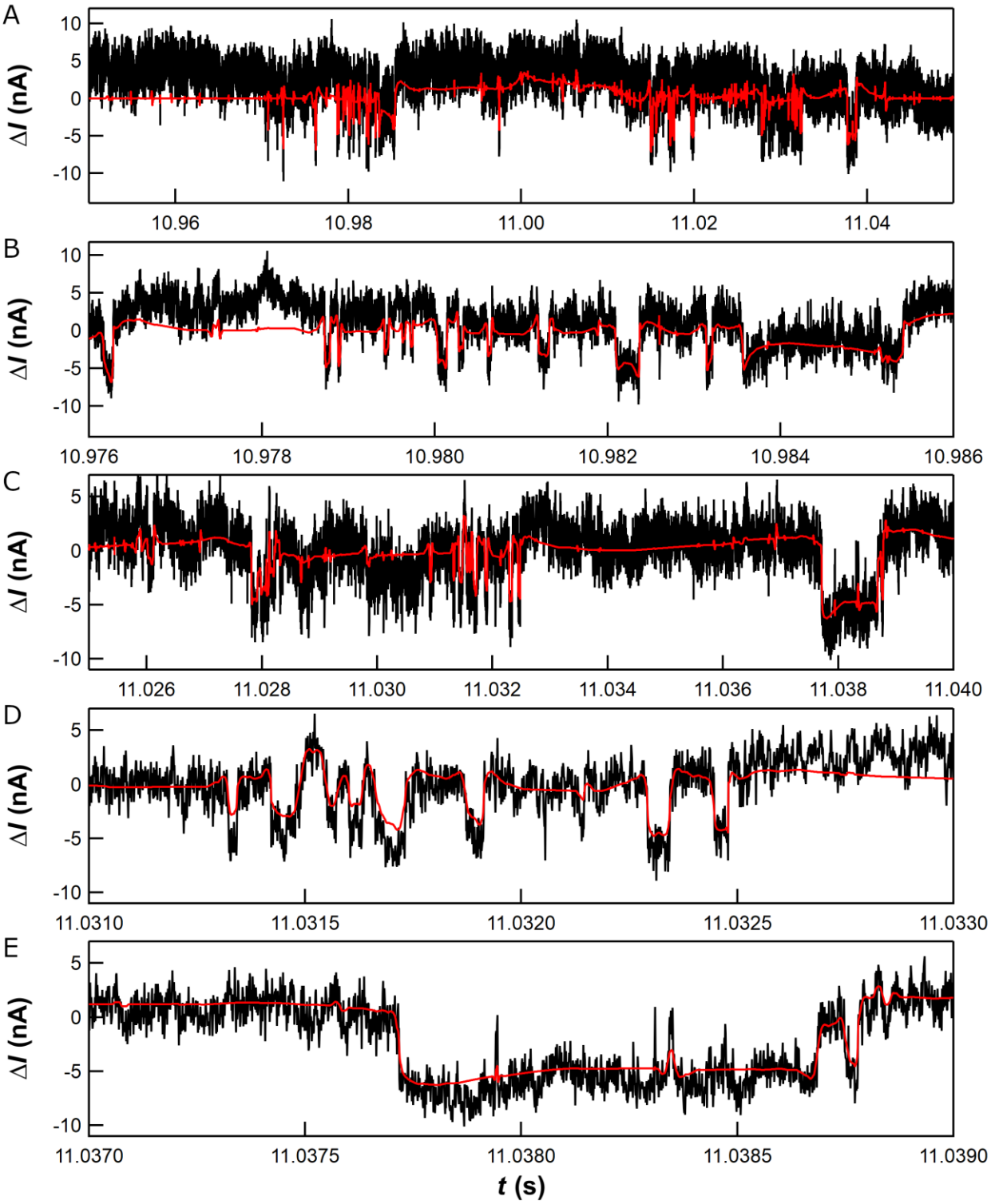

Figure 4.12: Examples of raw (black) and low-pass wavelet-denoised (red) signal from the SWCNT-FET, at a variety of timescales: (A) 100 ms, (B) 10 ms, (C) 15 ms, (D) 2 ms, and (E) 2 ms.

Generally, the amplitude and shape of the spikes and sharp transitions are preserved except when the spikes are less than ~100 μs in duration, as shown in Figure 4.12D, where the ~50-100 μs spikes which were originally rectangularly-shaped have been rounded at the corners and shrunk by ~20-50% in amplitude. In addition, some short-duration spikes are completely smoothed away by the wavelet denoising procedure, as can be seen in Figure 4.12E. Though the 1 ms rectangular wave which forms the overall shape of the spike is well-preserved, the spikes in the middle, returning to the baseline, are reduced by over 70% in the low-pass denoised signal.

Though the low-pass denoised channel smooths away the short spikes (< 100 μs in duration), the high-pass denoised channel does not. Figure 4.13A displays the same rectangular wave from Figure 4.12E, displaying the raw signal (black, top) alongside both the high-pass (blue, middle) and low-pass (red, bottom) denoised channels. The traces normally overlap but have been offset in the Y-axis for clarity. Though the low-pass channel smooths away > 90% of the amplitude of the ~10 μs spikes in the middle of the rectangular wave, the high-pass channel preserves the spike amplitudes entirely. Though the high-pass channel does not perform as much smoothing as the low-pass channel, it still reduces the low-amplitude noise fluctuations, as can be seen in the zoomed-in snapshot in Figure 4.13B, where the high-pass signal shows short segments of flat signal in between clusters of spikes or other oscillations. In addition, the slight low-frequency bump in the middle of the rectangular wave (at t = 11.0381 s) has been reduced by > 50%. Thus, the low-pass denoised channel performs well at smoothing the signal and preserving signal spikes longer than 100 μs, while the high-pass denoised channel captures the signal spikes and other features shorter than 100 μs.

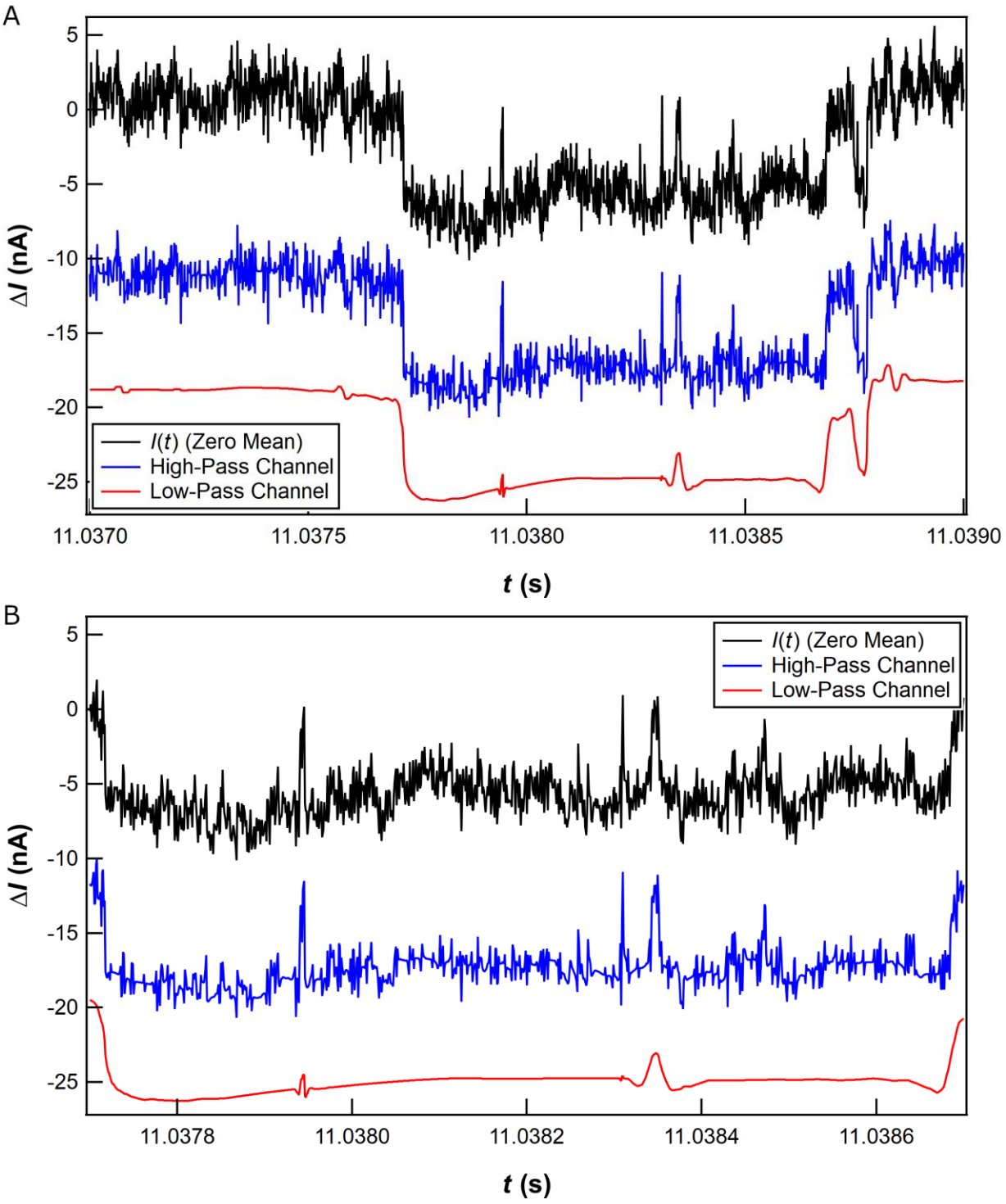

Figure 4.13: Example of the result of the high-pass (blue) and low-pass (red) wavelet denoising procedures, compared to the raw data (black), operating on the same signal as displayed in Figure 4.12E, shown on (A) 2 ms and (B) 1 ms timescales. The signal traces have been offset by -12 nA (high-pass) and -20 nA (low-pass) for clarity.

Figure 4.14 shows another example of the performance of the two denoised channels (high-pass in blue and low-pass in red, both offset from the raw signal in black) on a signal containing signal features ranging from 10 - 200 μs in duration. The output from the low-pass channel looks smooth and preserves the rectangular signal feature from 100-300 μs, but almost completely removes the ~10 μs spikes at t = 500 μs, and even halves the amplitude of the ~30 μs rectangular feature at t = 700 μs. By contrast, the high-pass channel preserves the ~10 μs features while reducing the amplitude of noise oscillations by ~50%, but the noise RMS remains 4 times greater than in the low-pass channel. The low-pass channel is optimized to preserve spikes and rectangular waves that are between 100 μs - 5 ms in duration, while the high-pass channel performs best on spikes shorter than 100 μs.

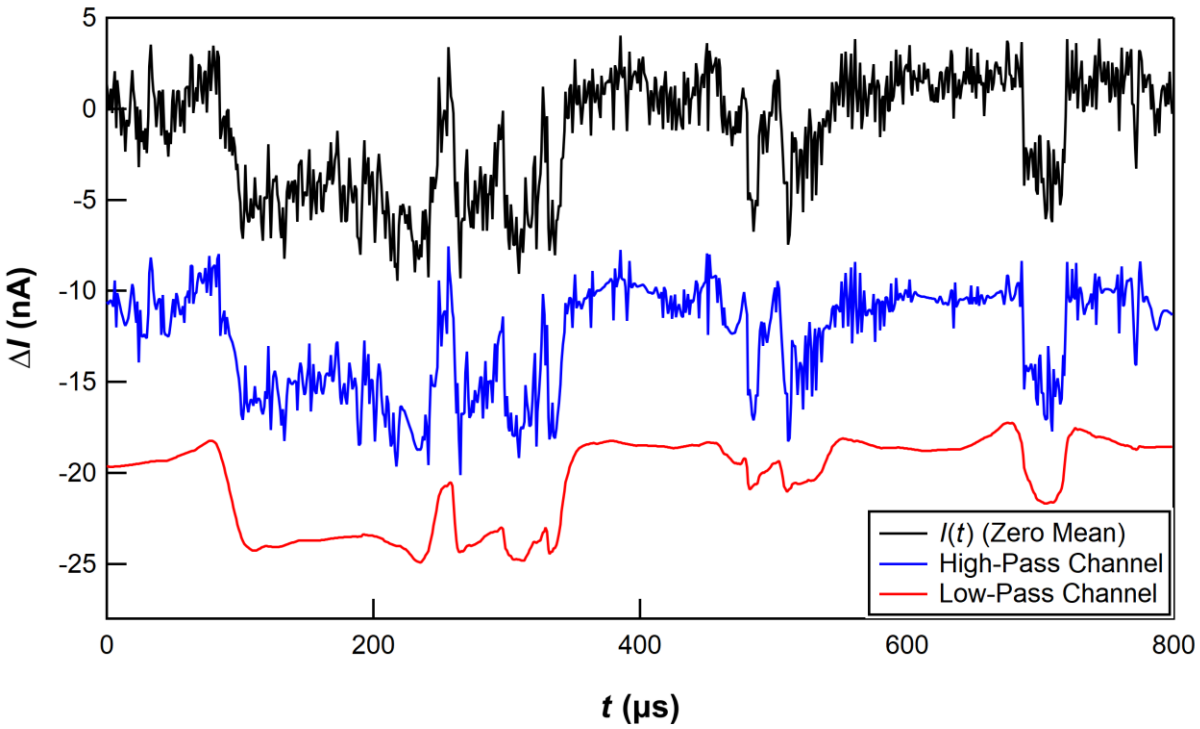

Figure 4.14: Example of the result of the high-pass (light blue) and low-pass (red) wavelet denoising procedures, compared to the raw data (black). The signal traces have been offset by increments of 10 nA for clarity.

4.4.3 Selecting the Optimal Wavelet Basis

The mother wavelet used as the basis function for wavelet denoising should match the signal of interest, which for typical enzymatic activity is composed of sharp spikes and rectangular jumps on an otherwise flat baseline. A close match between the mother wavelet and the signal allows the wavelet transform to encode the signal in fewer, larger coefficients that are preserved by thresholding, resulting in less distortion. From the list of known mother wavelets, the leading candidates for the mother wavelet are (Figure 4.15): the Haar wavelet, Coiflet wavelet family, and the biorthogonal wavelets (particularly the biorthogonal 1.3 and 1.5 wavelets). The Coiflet family is best at identifying sharp spikes, while the other wavelets mentioned are ideal for identifying edges in the signal. Most other wavelet families (not illustrated here) have shapes that resemble ripples or noise rather than discrete jumps (149, 169).

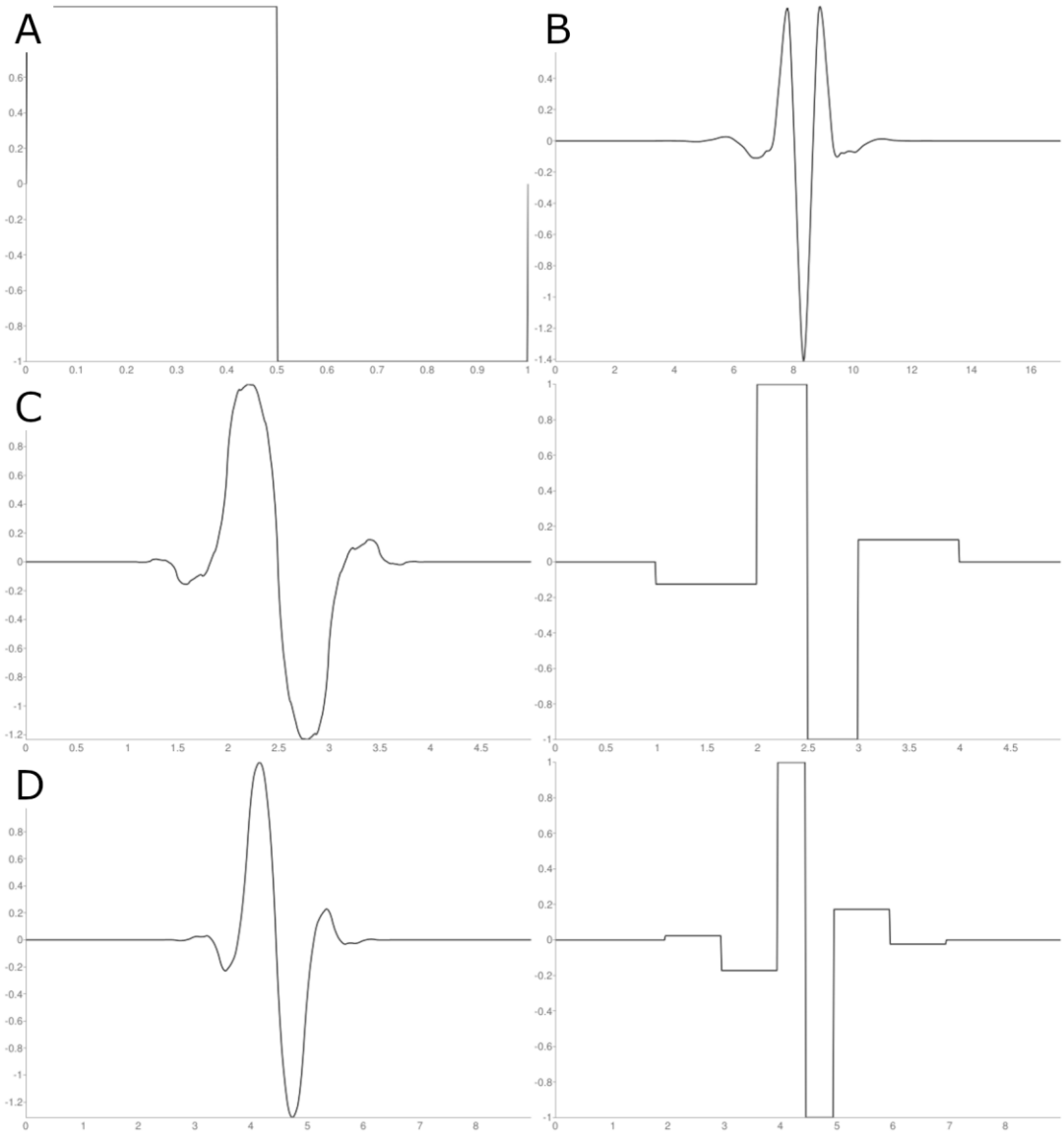

Figure 4.15: Mother wavelet candidates for denoising the SWCNT-FET signal: (A) Haar, (B) Coiflet 3, (C) biorthogonal 1.3 (D) biorthogonal 1.5. Note that the two biorthogonal functions have two mother wavelets: one for decomposition, the other for reconstruction.

Evaluation of the performance of these mother wavelets was performed using data obtained from Taq DNA polymerase enzymatic activity. The denoising algorithm was run using each candidate mother wavelet, with all other parameters remaining the same, over the same set of data. The optimal mother wavelet should result in the flattest baseline and the sharpest

edges. Figure 4.16 shows histograms of the resulting denoised signals using the four candidate wavelet bases, with the histogram of the raw data in black for comparison. The histograms mostly overlap, especially on a log scale (Figure 4.16A). However, a linear scale (Figure 4.16B) reveals that Haar wavelet basis produces a 45% higher peak near 0 nA (the baseline level) than the next-highest peak (corresponding to the biorthogonal 1.3 wavelet). This means that the Haar wavelet basis is the best at bringing the signal in the baseline to the 0 nA level, and thus results in the flattest baseline.

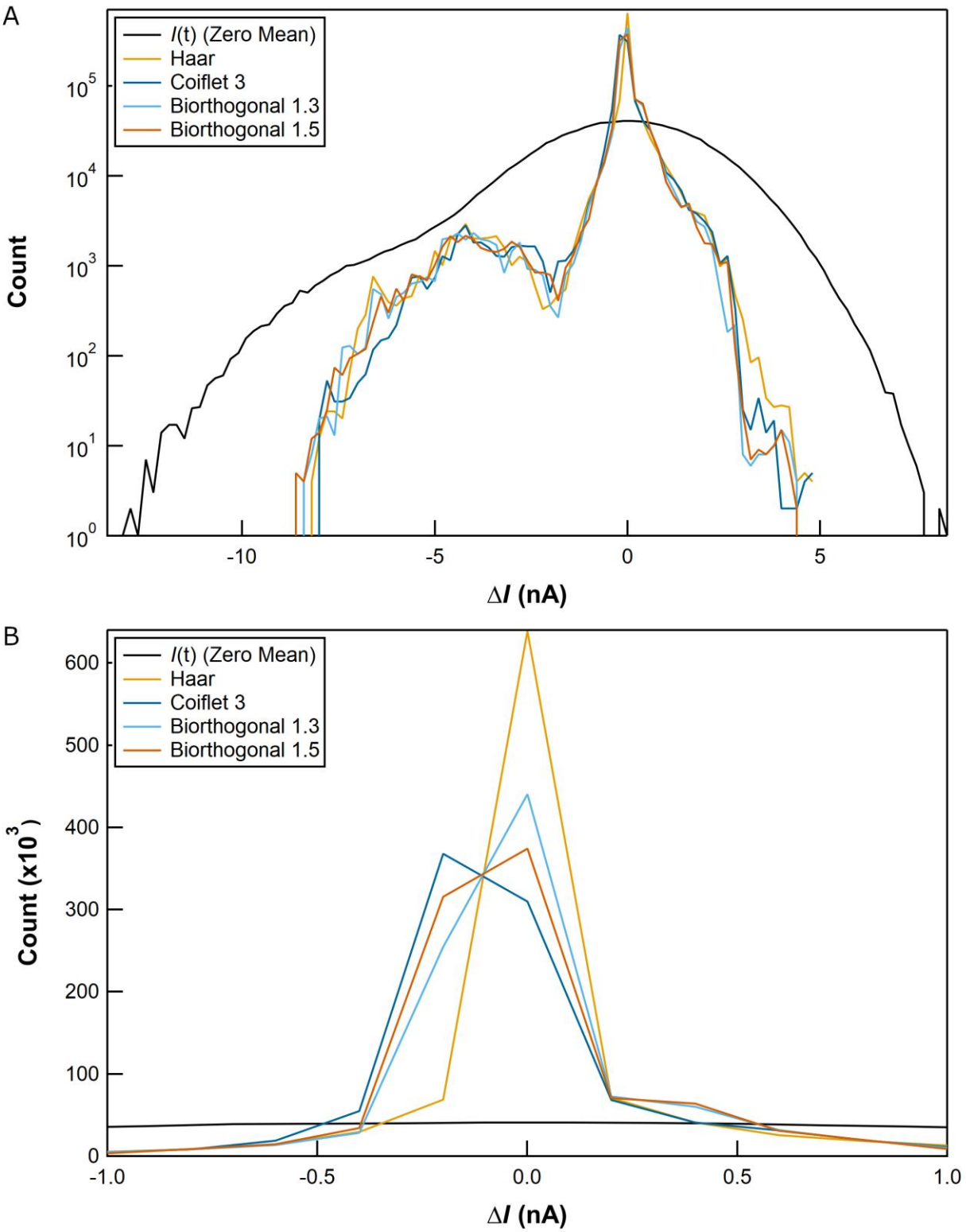

Figure 4.16: Histograms of denoised SWCNT-FET signal using various wavelet bases, shown in both log (A) and linear (B) scale.

In addition, wavelet denoising using the Haar wavelet basis results in smaller artifacts than with the other three wavelet bases. Figure 4.17 shows (in black) an idealized, noise-free version of the rectangular 2-state signal that commonly occurs in SWCNT-FET recordings of biomolecules. The result of applying the wavelet denoising procedure to this signal, using the four candidate wavelet bases functions, is overlaid on top of the original signal. Generally, the denoising procedure produces similar artifacts regardless of wavelet basis: the baseline (at 0 nA) develops a hump at the location of the rectangular discontinuity, and the position of the second state (originally at -1 nA) is shifted toward the original baseline. The magnitude of the hump and the resulting shift is dependent on both the magnitude and duration of the original rectangular discontinuity. For this particular example (a 10 ms rectangular spike of amplitude 1 nA in a 200 ms window), the hump is ~0.28 nA tall for the Coiflet 3 and biorthogonal wavelets. However, the Coiflet 3 wavelet and the two biorthogonal wavelets produce an additional artifact: ripples in the originally-flat baseline whose duration is ~10 times the duration of the original rectangular signal. By contrast, the Haar wavelet basis produces no ripples beyond the hump, and the magnitude of the hump is reduced (~0.26 nA instead of ~0.28 nA), such that the position of the second state is closer to the original. Thus, the Haar wavelet function is the best basis to use for denoising SWCNT-FET signals.

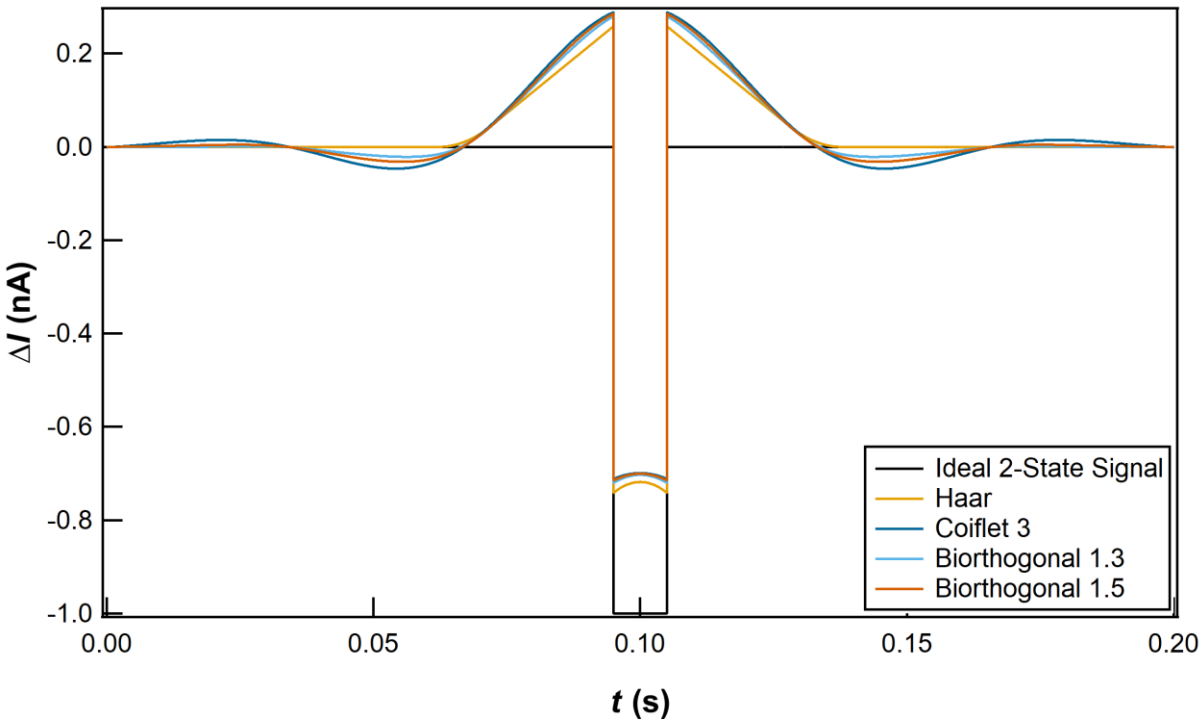

Figure 4.17: Comparison of the distortion of an idealized 2-state signal (black) produced by the wavelet denoising algorithm, using four wavelet bases.

4.4.4 Analysis of the Denoised Signal

One primary goal in denoising the SWCNT-FET biosensor signal is to facilitate separation of the signal into distinct states, identified by finding peaks in the histogram of the signal and fitting those peaks to gaussian functions. There are two aspects of state separation: flattening the baseline (removing low-frequency fluctuations) and reducing the noise band or RMS width (removing high-frequency fluctuations). Both aspects serve to reduce the widths of the peaks in the histogram, increasing the likelihood of finding local minima (valleys) between peaks that denotes a boundary between states. The standard analysis procedure applied to a raw $I(t)$ signal consists of: 1) a wavelet denoising algorithm, and 2) an automated spike-detection and state-finding algorithm on each denoised channel.

Figure 4.18A and B compares the histograms of the raw and denoised signals for two different 1 s segments. In Figure 4.18A, the histogram of the original signal shows a wide and round primary peak (corresponding to the baseline state) with a significant shoulder on the left side, suggesting the presence of the second state. The histogram of the high-pass denoised signal shows that the second state now has its own peak, separated from the first peak by a valley, and the widths of the two peaks have been reduced. The histogram of the low-pass denoised signal shows peaks with even smaller widths, and the presence of a slight bump on the far left of the secondary peak may suggest the presence of a third peak. In addition, the position of the second peak has shifted right (toward the first state). The width of the primary peak is 10 nA in the original signal, 6 nA in the high-pass denoised signal, and 2 nA in the low-pass denoised signal. In the histogram, the widths of the peaks are correlated with the amplitude of the noise in the signal, so the narrower peak widths in the denoised signals indicate a lower noise amplitude.

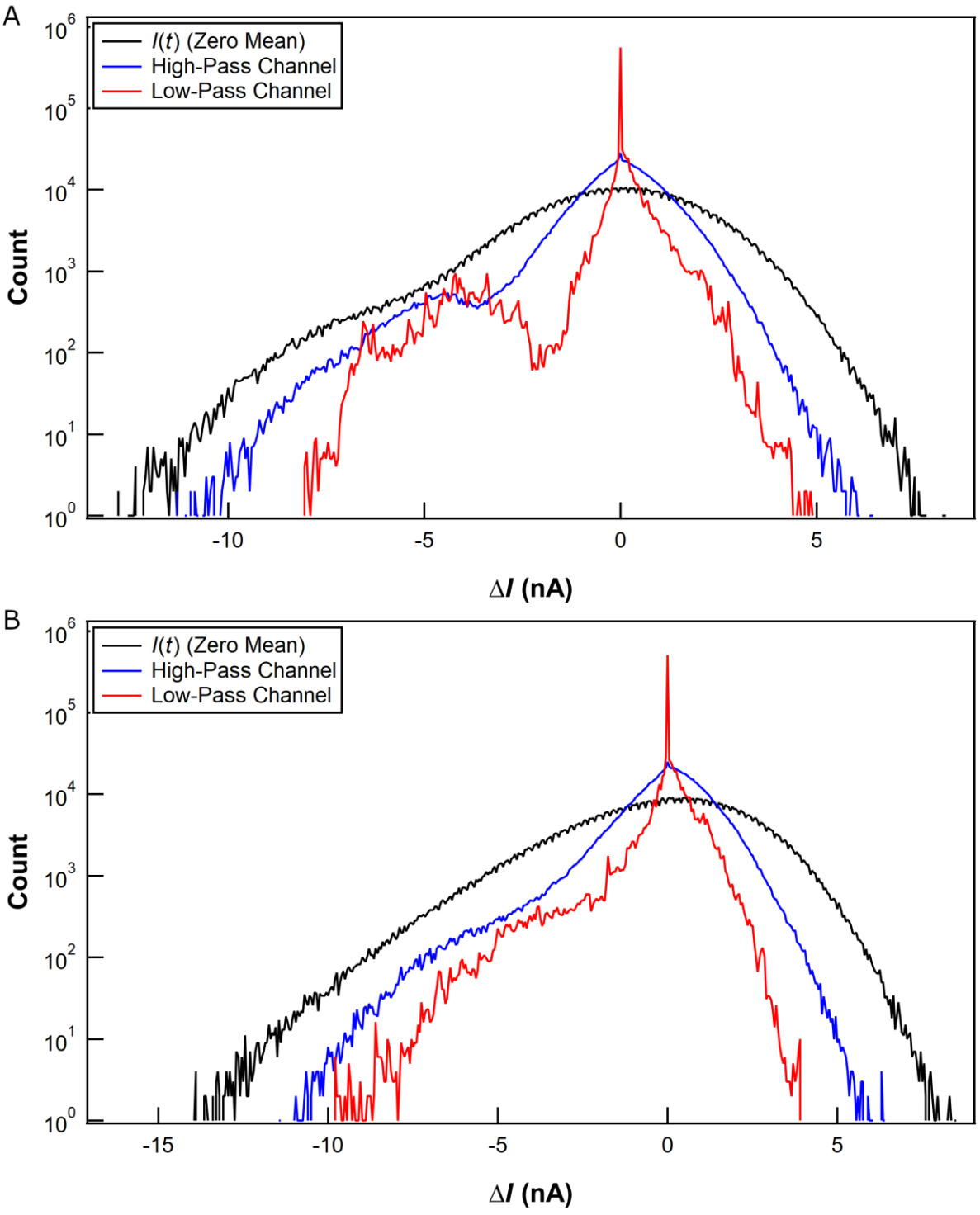

Figure 4.18: Histograms of the raw signal (black), high-pass wavelet-denoised (blue), and low-pass wavelet-denoised (red) signals, for two different 1 s segments of signal.

In Figure 4.18B, the corresponding signal does not have many samples in the second state. Thus, the amplitude of the second peak is low, and the histogram of the original signal appears as a single, smooth, wide hump because the second peak is hidden by noise and low-frequency fluctuations. The histogram of the high-pass channel shows a reduced width in the primary peak, revealing a shoulder corresponding to the second state. In the histogram of the low-pass channel, the width of the primary peak is further reduced, making the shoulder even more distinct, although the position of the shoulder is also shifted toward the baseline state. Though the second state does not appear as a distinct peak in this example, peak fitting methods can still determine a position for the second state. This example demonstrates the ability of the wavelet denoising procedure to separate the signal into distinct states, although the low-pass denoised channel sometimes shifts the position of any non-baseline states toward the baseline position.

When the raw $I(t)$ contains enough points in non-baseline states, the histogram of the raw $I(t)$ can be modeled as a sum of gaussian peaks, with each gaussian denoting an independent state, and the positions and widths of the gaussians can be determined by fitting. The wavelet denoising procedure alters the shapes of the resulting histograms by reducing the width of the peaks and sometimes shifting the positions of the secondary peaks toward the baseline (the highest peak). The denoising also makes the peaks non-gaussian and asymmetric – in particular, the peaks become too narrow near the tip to be accurately modeled as a gaussian. The distortion of the peaks away from their originally-gaussian shape is due, in part, to aggressive smoothing (which concentrates more data points near the peak centers) and to the artifacts introduced by the low-frequency suppression in the wavelet

denoising procedure (as discussed in Section 4.4.3 and illustrated in Figure 4.17). However, it is currently unknown whether additional factors contribute to the asymmetry or non-gaussian nature of the denoised peaks. Determining any additional contributions will require further investigation.

4.5 Comparison of Denoising Approaches

As previously described, the SWCNT-FET signal contains several types of noise that are intermixed with the biomolecule signal in different frequency ranges, and each frequency range can be handled with different denoising approaches. For instance, the wavelet denoising procedure described above removes low-frequency fluctuations by zeroing the approximation coefficients and the coefficients in the low-frequency scales, effectively creating a high-pass filter with a cutoff at ~60 Hz. By contrast, in previous work (1-3, 25), low-frequency fluctuations were removed by fitting an interpolated spline to a decimated $I(t)$ and then subtracting the spline from the raw $I(t)$. Figure 4.19 compares a 2 s segment of the raw $I(t)$ to the previously-used interpolating spline subtraction method, along with the two wavelet denoising channels and an FIR high-pass filter, with the signal traces offset for clarity. The differences in the RMS of the various signals (as illustrated by the widths of the signal trace) are due to differences in high-frequency filtering. At this scale, the denoising methods seem to perform roughly equally well in removing the ~2-5 Hz oscillations in the raw $I(t)$, since the centers of the signal traces remain flat. A 200 ms zoomed-in snapshot (Figure 4.20) shows that the FIR highpass filter completely flattens the ~40 ms bump at $t \sim$ 11 s, while the wavelet denoising methods are the worst at flattening this bump.

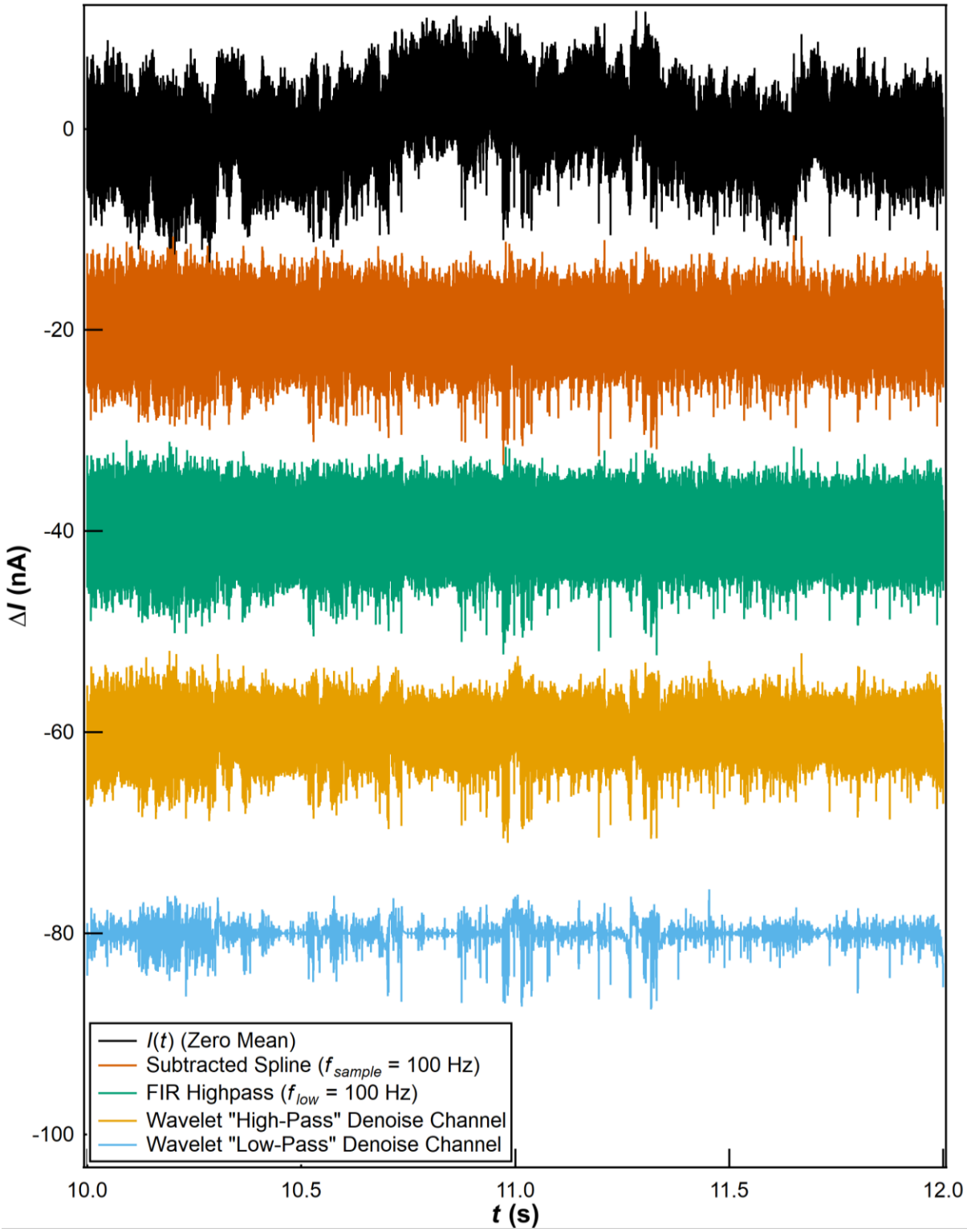

Figure 4.19: Comparison of 2 s segments of the raw $I(t)$ (black, top) with the output from: subtracted spline low-pass filtering (orange, second), FIR bandpass filtering (green, third), wavelet "high-pass" denoising (yellow, fourth), and wavelet "low-pass" denoising (blue, bottom). The signal traces are offset by 20 nA increments for clarity.

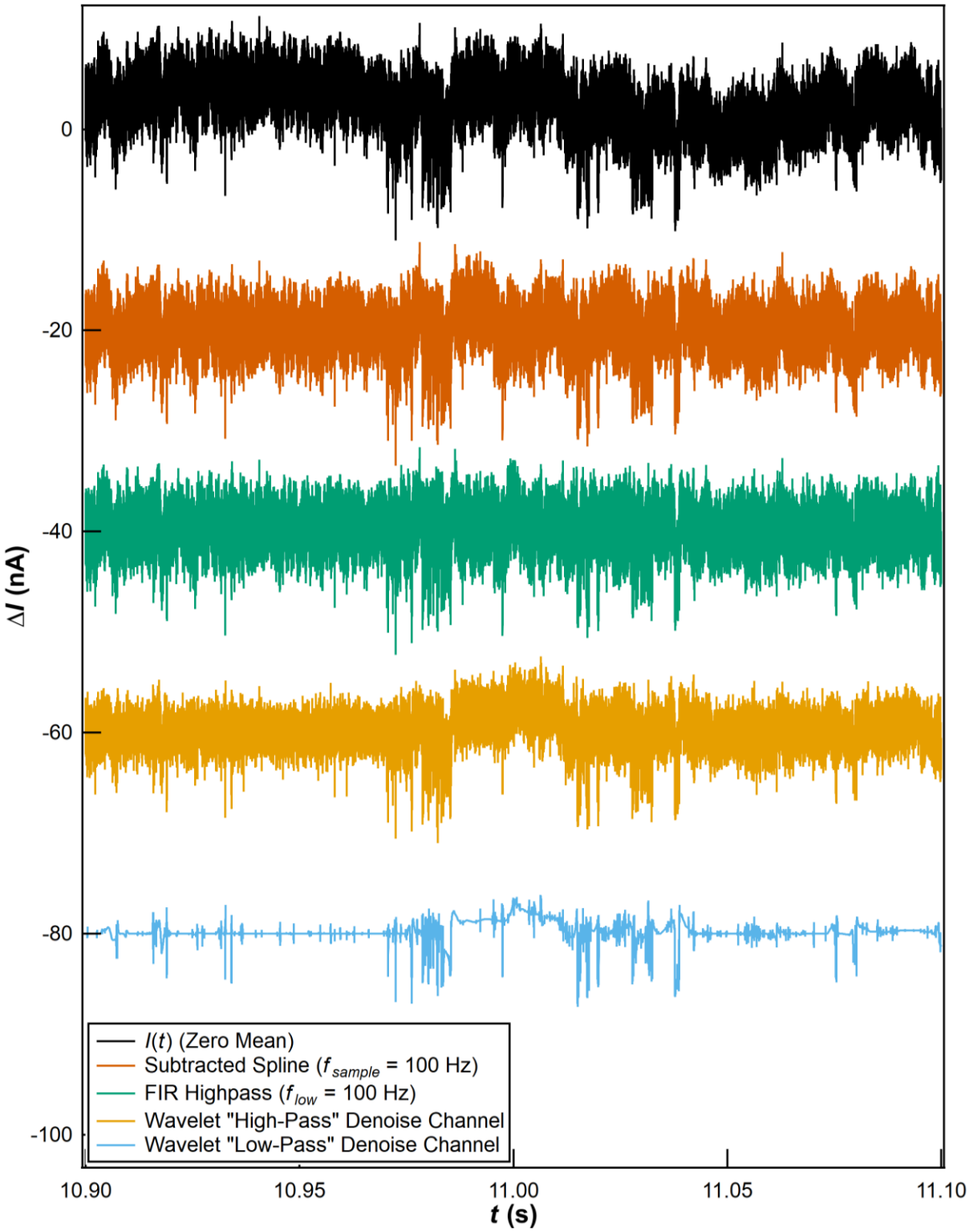

Figure 4.20: Comparison of 200 ms segments of the raw $I(t)$ (black, top) with the output from: subtracted spline low-pass filtering (orange, second), FIR bandpass filtering (green, third), wavelet "high-pass" denoising (yellow, fourth), and wavelet "low-pass" denoising (blue, bottom). The signal traces are offset by 20 nA increments.

The effectiveness of low-frequency denoising can be quantified using histograms of the signal, as shown in Figure 4.21. The histogram of the raw $I(t)$ is a wide, asymmetric hump whose peak is at ~2 nA. The histograms of the denoised signals show peaks at 0 nA, showing that the baseline has been centered at 0, and the peaks are roughly symmetric around the peak centers. In addition, the widths of the peaks are reduced by ~30% relative to the peak for the raw $I(t)$, which allows a small shoulder corresponding to the second state to appear on the left. The peak of the wavelet "low-pass" denoising is further reduced, allowing the peak of the second state to appear, because the filter also reduces high-frequency noise.

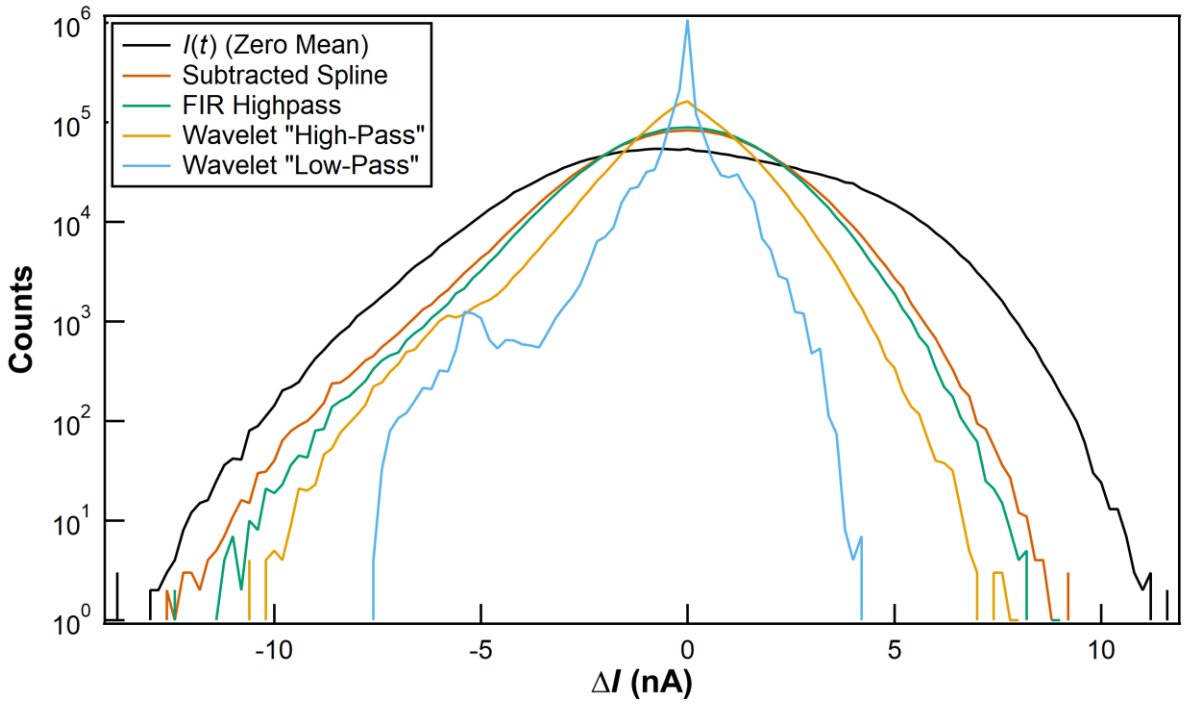

Figure 4.21: Histograms of the 2 s segments of signal illustrated in Figure 4.19.

Removing the high-frequency oscillations can be achieved through several methods, including: wavelet denoising, FIR filtering, median filtering, total variation denoising, and

NoRSE filtering. The following 6 figures will demonstrate the effects of the various high-frequency denoising methods on the same 8 ms segment of signal, with the same spikes labeled with lowercase letters in each figure for clarity. Figure 4.22 shows the result of the two wavelet denoising channels. Note that both channels completely preserve the two ~100 µs duration spikes (at t ~ 2.5 ms, labeled a) and mostly preserve the ~50 µs spike at t ~3.5 ms (labeled c). The low-pass channel reduces the amplitude of the ~50 µs spike by ~20% but also substantially reduces the baseline RMS (~0.3 nA instead of ~1.8 nA for the raw $I(t)$), while the high-pass channel preserves the spikes entirely (including the ~20 µs spike labeled b) and reduces the baseline RMS to ~0.8 nA. The high-pass channel also causes the middle of the ~4 ms rectangular waveform to bulge upward, reducing the amplitude of the wave.

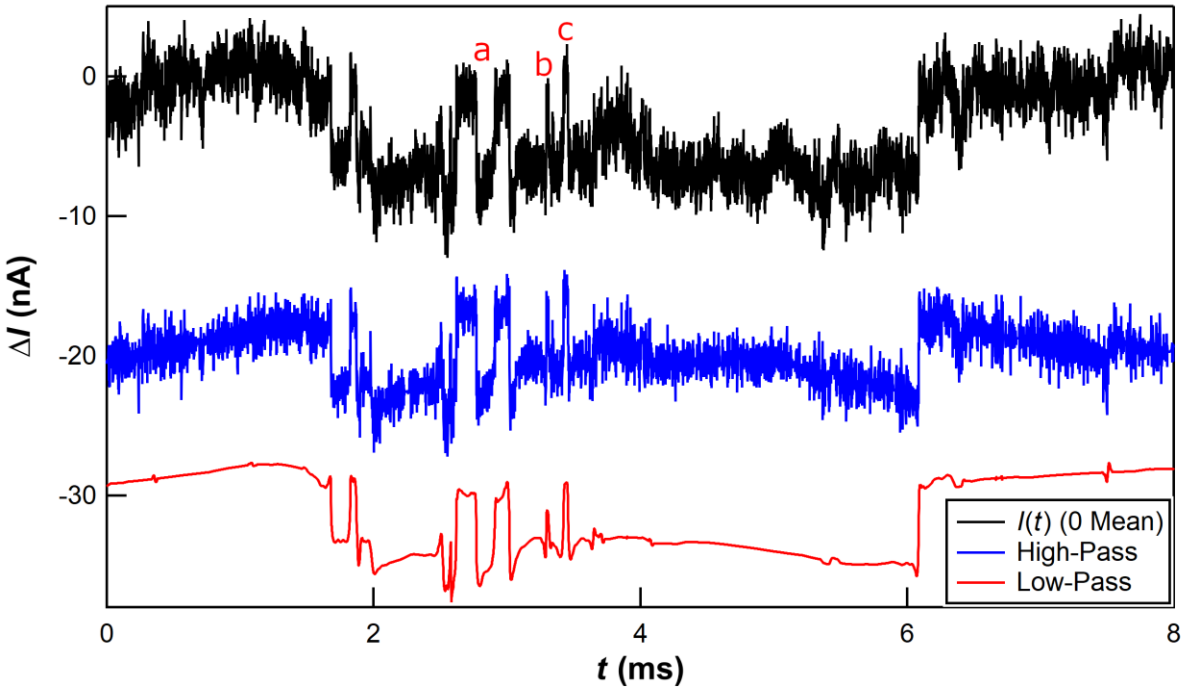

Figure 4.22: An example SWCNT-FET signal (black) passed through the "high-pass" (blue) and "low-pass" (red) wavelet denoising channels. The signal traces are offset by the tick values for clarity.

FIR filters can be utilized, but their frequency-based design means that sharp spikes and transitions are usually rounded or smoothed. Figure 4.23 shows examples of FIR bandpass filters with varying low-pass cutoff frequencies (f_{high}). The filters with a high f_{high} (such as f_{high} = 40 kHz, orange) reduce the baseline RMS from 1.8 nA to 1.2 nA while preserving the shapes and positions of the spikes labeled a, b, and c. Filter with f_{high} < 40 kHz reduce the amplitude of spike b, while filters with f_{high} < 10 kHz) reduce the amplitude of (or even completely smooth away) spikes a and c. In addition, the baseline RMS for filters with f_{high} < 10 kHz, which is ~ 1.1 nA, does not differ much from the baseline RMS for filters with f_{high} ~ 80 kHz (1.3 nA).

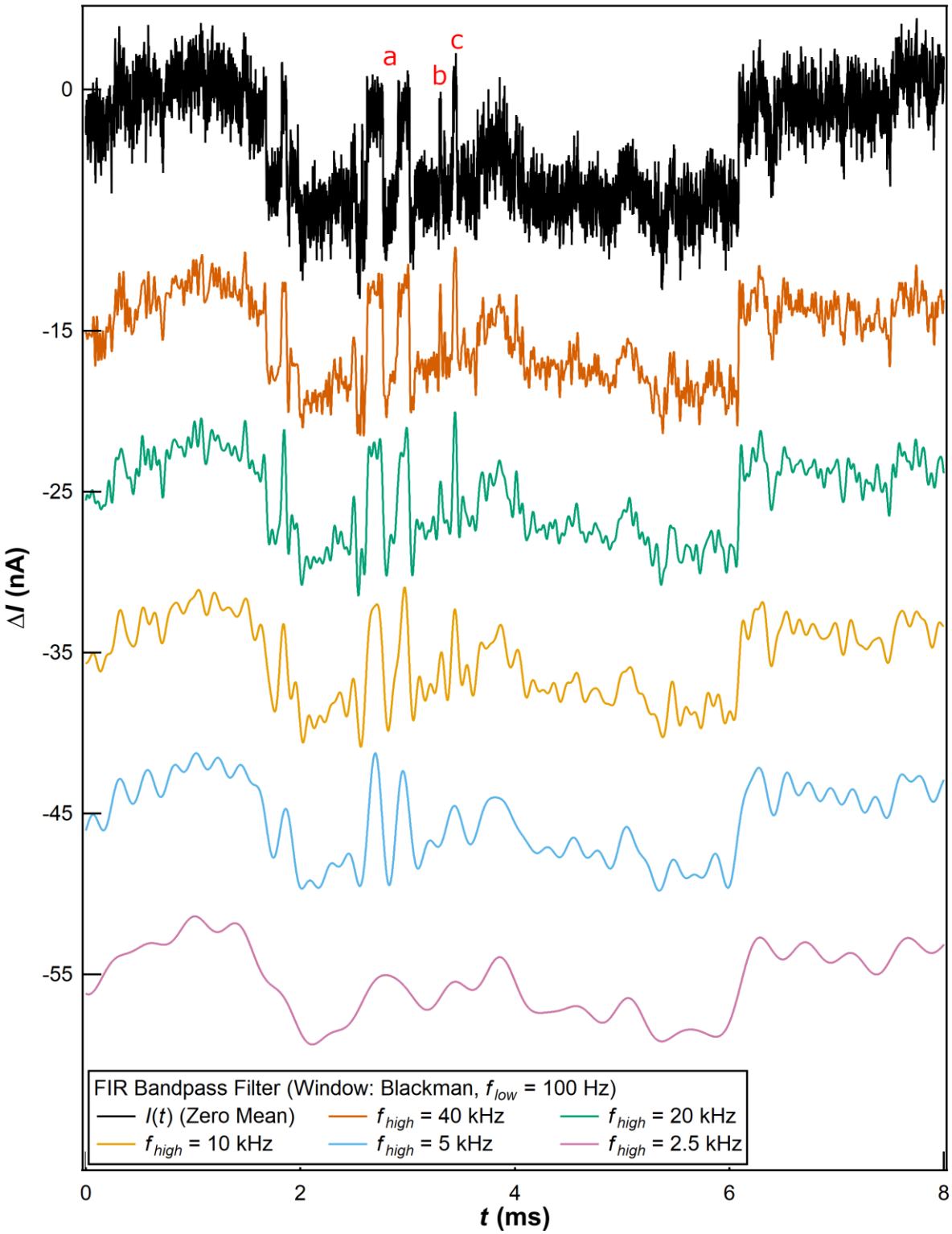

Figure 4.23: An example SWCNT-FET signal passed through FIR bandpass filters with constant f_{low} and various f_{high} values. The signal traces are offset by the tick values for clarity.

Median filters are particularly suited for smoothing high-frequency noise and removing spikes in a signal (including salt-and-pepper noise in images). Unfortunately, median filters operate on a fixed characteristic timescale or frequency, so features shorter than the characteristic timescale are removed or, at least, are reduced in amplitude. As shown in Figure 4.24, a median filter with kernel size 257 or larger removes oscillations in the baseline (RMS 1.2 nA) but also removes the significant spikes in the signal, while kernel sizes of 65 (baseline RMS 1.3 nA) or 129 preserve the two spikes labeled a but suppress spike c. Only the filter with kernel size 17 (baseline RMS 1.4 nA) preserves spike b.

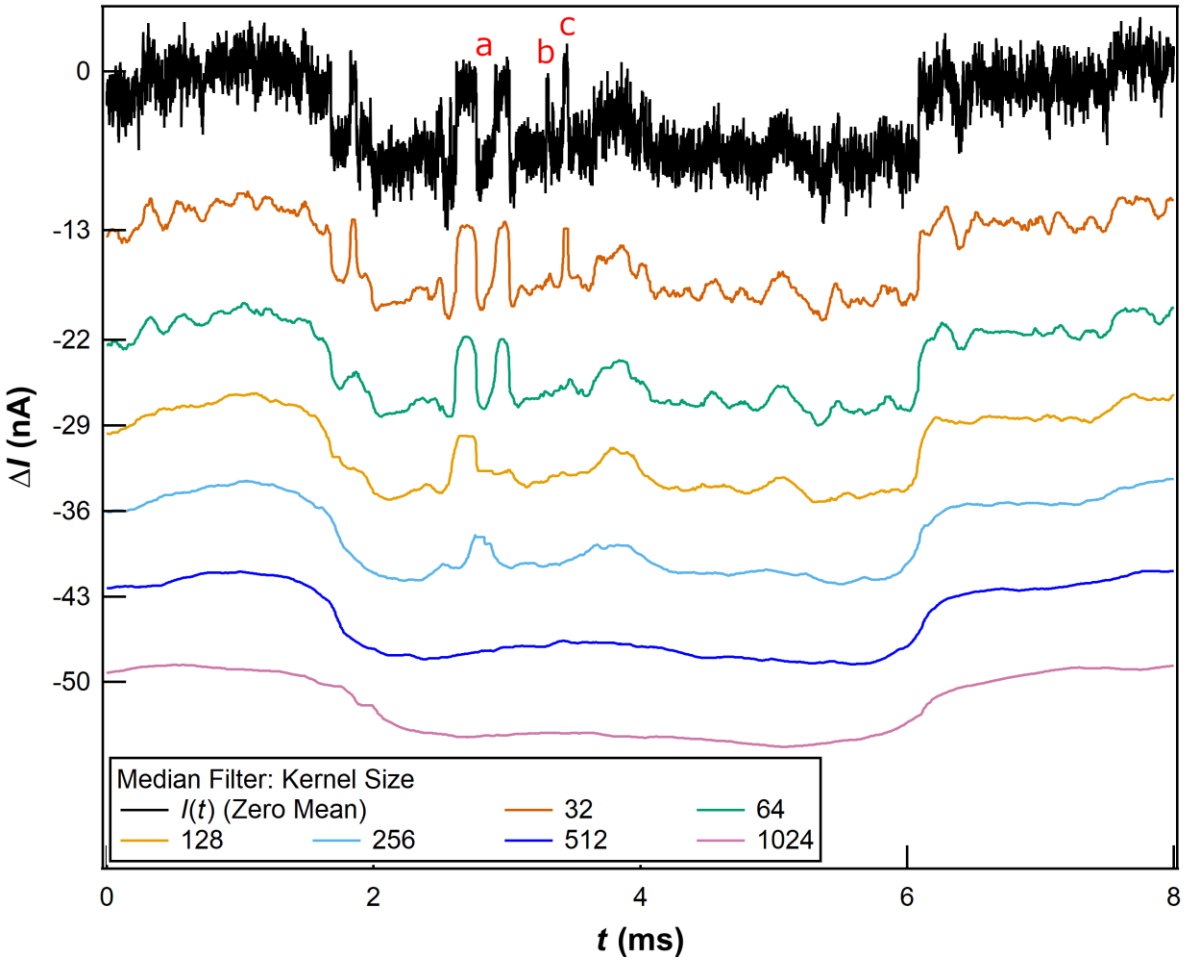

Figure 4.24: An example SWCNT-FET signal passed through median filters with varying kernel sizes. The signal traces are offset by the tick values for clarity.

To first order, an idealized signal from biomolecules measured by SWCNT-FET sensors can be modeled by finite-state systems. A noise-free signal without any other degrees of freedom consists of sharp jumps between two or more consistent states. Two denoising procedures which preserve sharp jumps while suppressing small-amplitude oscillations are total variation (TV) denoising and the NoRSE filter. Total variation denoising tries to minimize the total "distance" traveled by a signal, making the output block-like with regions of flat, discrete levels and sharp transitions between them (146, 183). Figure 4.25 displays the output of the total variation denoising using different values of the parameter λ. For $\lambda =$

1×10^8, the baseline RMS is 1.4 nA, and the denoising procedure preserves spikes a and c and mostly preserves spike b. For $\lambda = 1 \times 10^7$, the baseline RMS is 1.2 nA, and the denoising procedure preserves spikes a but mostly smooths away spikes b and c. Note that at this higher λ, the signal RMS is not substantially different than that from lower values of λ.

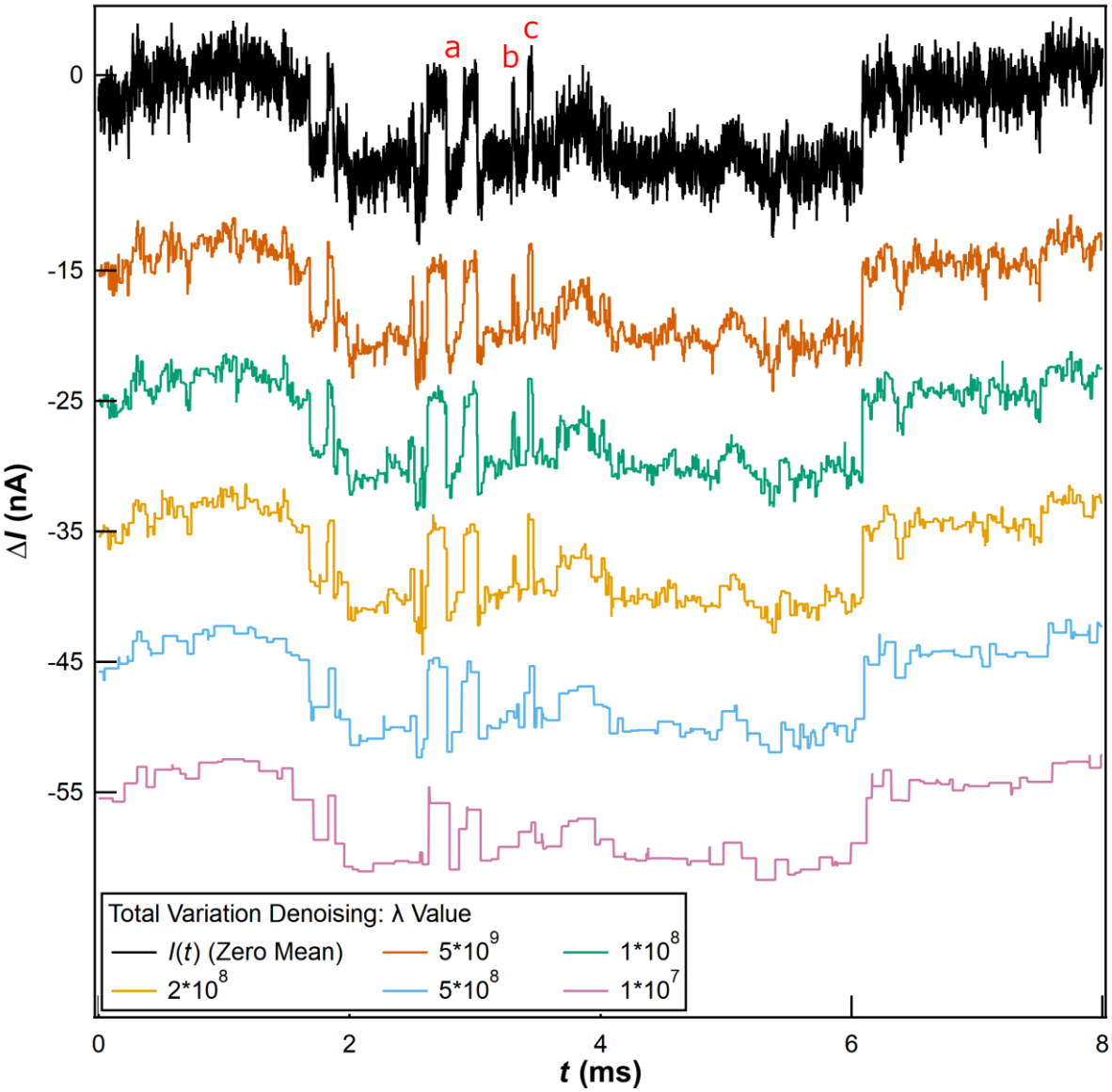

Figure 4.25: An example SWCNT-FET signal passed through total variation denoising with varying values of the parameter λ. The signal traces are offset by the tick values for clarity.

The NoRSE (Noise Reduction and State Evaluator) algorithm (75) looks forward and backward in time to find sharp transitions in the signal, smoothing the signal everywhere except at transitions. The algorithm repeats this forward-backward search as many times as the number of filters in the filter bank. Here, the NoRSE algorithm has been implemented with a bank of 5 filters, each larger than the previous by a multiple of 2. Figure 4.26 displays the output of the NoRSE filter applied to the raw $I(t)$ (black) with varying filter sizes. With smaller filter sizes (filter sizes [2, 4, 8, 16, 32]), the NoRSE filter preserves both significant signal spikes as well as noise oscillations, and the baseline RMS is 1.5 nA. For filter sizes [32, 64, 128, 256, 512] (which produces a baseline RMS 1.3 nA) or larger, the filter reduces the amplitude of spikes b and c. For even larger filter sizes (greater than [512, 1024, 2048, 4096, 8192]), the filter generates sharp spikes as artifacts, compromising the effectiveness of the denoising.

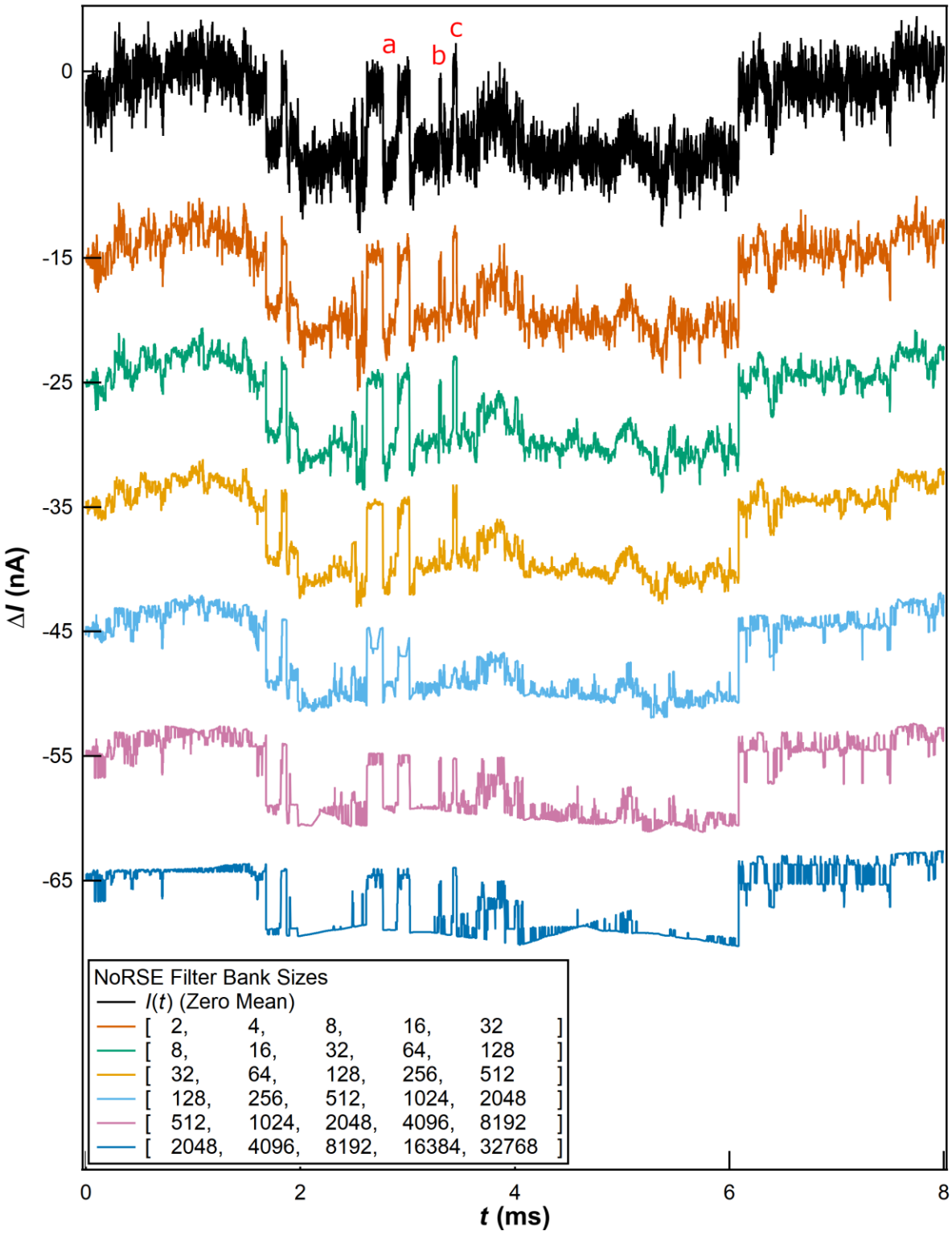

Figure 4.26: An example SWCNT-FET signal passed through the NoRSE filter with varying sizes of filter banks. The signal traces are offset by the tick values for clarity.

Another recent development in signal denoising is adaptive filtering, in which the filter automatically adjusts its own parameters to optimize noise removal. Adaptive versions of FIR/IIR filters (184), median filters (185), and total variation denoising (186, 187) do exist, and utilizing such adaptive filters could, in theory, adjust the filter parameters around short-duration spikes to preserve them in the denoised signal. However, all adaptive filters require an independent channel to capture noise, which is correlated to the noise in the signal channel but also independent of and uncorrelated with the signal itself. Unfortunately, for SWCNT-FET biosensors, it is extremely difficult (if not impossible) to create an independent channel whose noise correlates with the original channel, because the dominant noise sources are the Schottky barriers and atomic impurities at the junction where the SWCNT and the electrode meet, and the characteristics of each particular junction are highly variable and difficult to control. Even when another electrode is connected to the sensing SWCNT-FET as a separate channel, the noise measured from the new electrode will have different characteristics (RMS, corner frequency, etc). If fabrication of SWCNT-electrode contacts can be reliably controlled, adaptive filtering can be used for denoising SWCNT-FET signals.

A comparison of the output from both high- and low-pass wavelet denoising channels, median filter, total variation denoising, and NoRSE filter is shown in Figure 4.27, with parameters chosen to preserve ~50 μs spikes. The low-pass wavelet denoising algorithm reduces the baseline RMS the most while still preserving the amplitude of spikes a, and also does well at flattening round bumps in the signal. The high-pass wavelet denoising procedure accurately preserves all spikes shorter than 200 μs. Total variation denoising accurately preserves sharp transitions and spikes longer than 100 μs but filters out shorter

spikes. The NoRSE filter accurately preserves spikes shorter than 200 µs but creates spikes as artifacts, and still only reduces baseline RMS as much as the wavelet high-pass method.

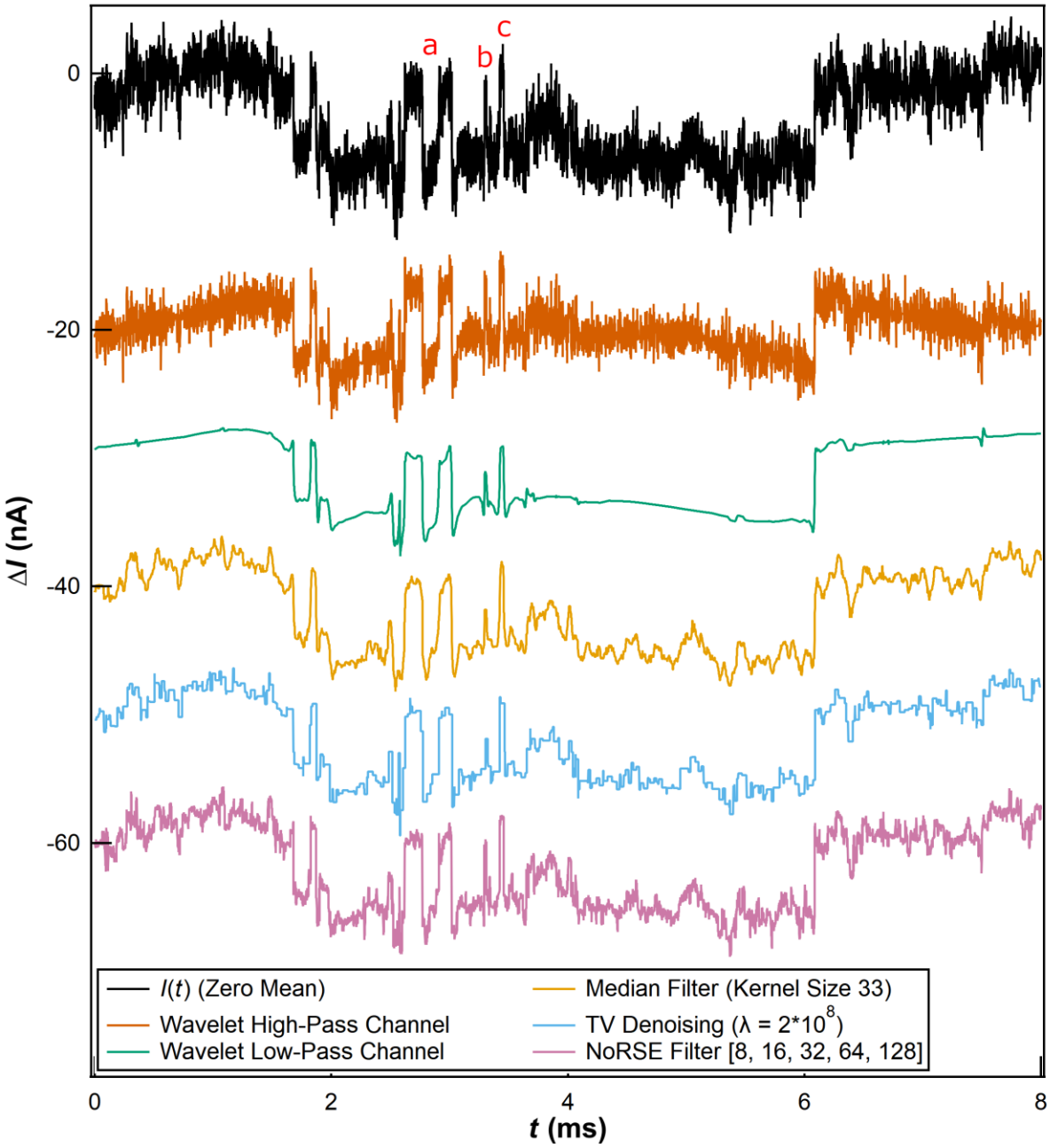

Figure 4.27: Comparison of denoising methods. The parameters for each method are: wavelet high- and low-pass channels as described in Table 4.1, median filter kernel size = 33, total variation denoising $\lambda = 2 \times 10^8$, NoRSE filter bank = [8, 16, 32, 64, 128].

The wavelet algorithm does have some weaknesses. In particular, the low-pass wavelet denoising parameters used here reduce the amplitude of the overall rectangular wave by ~20% as compared to the other methods. Also, the low-pass wavelet denoising smooths away the ~20 μs spikes located on the plot between 6 and 8 ms, probably because the amplitudes of the spikes were smaller than the threshold at the relevant scale, while the NoRSE filter manages to preserve the spike almost entirely and total variation denoising reduces the spike amplitudes by ~50%. Though the high-pass wavelet denoising channel is able to preserve short spikes, it distorts the shape of rectangular waves with a duration longer than 1 ms, which does not occur with the other denoising methods.

The denoising methods described here vary in computational cost. With the parameters required to achieve the outputs as demonstrated in the figure, the computation time required to complete the denoising is least for total variation denoising, followed by median filtering, wavelet denoising, and finally NoRSE filtering. The computation time for both median and NoRSE filtering depends on the parameters chosen: as the kernel size or filter size increases, the computation time increases. By contrast, wavelet denoising carries an approximately-constant computational cost for a given signal, regardless of the parameters, but requires the most memory to compute due to the large number of UWT coefficients.

The most direct analogue to SWCNT-FET measurements of biomolecules are found in the fluorescence community. Algorithms that were developed to look for a multi-state output in a signal, such as vbFRET (188), HaMMy (189), and STaSI (190), could also be applied to the SWCNT-FET signal to remove baseline fluctuations and identify state transitions. When

these algorithms were applied to the SWCNT-FET signal, the algorithms performed similarly to wavelet denoising and a simple thresholding algorithm (detailed in chapter 5) in identifying transition locations, yet required orders of magnitude more computation time to complete. The primary obstacles for these FRET-based algorithms were the low signal-to-noise ratio and the fluctuating baseline of the SWCNT-FET signal, which are not typical problems encountered in FRET experiments.

Another analogue to SWCNT-FET denoising is the denoising of electroencephalograms (EEG) (170, 191) or electrocardiograms (ECG) (155, 156, 184, 186). EEGs are recordings of brain electrical activity, while ECGs are recordings of heart electrical activity. Both types of measurements consist of sharp spikes on a fluctuating or drifting baseline in the presence of high-frequency instrument noise and artifacts from other nerve activity. In these fields, wavelet denoising has become an increasingly popular technique. The wavelet transform handles EEG and ECG spikes particularly well, especially because some wavelet bases (like the Daubechies, symlet, or Coiflet families) match the general shape of the spikes. As a result, the significant signal information can be captured in fewer, high-amplitude coefficients, making the denoising particularly efficient.

Alternative methods in EEG and ECG denoising often use adaptive filters to remove the high-frequency noise and some sort of regression to remove the low-frequency drift (141, 155, 186). Some algorithms developed specifically for EEG or ECG denoising rely on models of known characteristics of the spikes. Also, since EEG and ECG measurements often use multiple electrodes, multi-channel denoising and decorrelation algorithms can be used,

including blind source separation (BSS) (142, 192), SCADS (193), and independent component analysis (ICA) (194, 195), among others. Such multi-channel denoising methods cannot apply to SWCNT-FET signals, which are limited to single channels.

4.6 Summary

This work applies wavelet-based denoising to signals from SWCNT-FET sensors and demonstrates the effectiveness of the denoising procedure for removing 1/f noise from the SWCNT-FET while preserving the spikes and rectangular transitions from biomolecule activity. The dyadic scaling of the wavelet transform makes it ideal for decorrelating 1/f noise, and the multi-scale nature of wavelet denoising provides flexibility to optimize the denoising separately in the low-, mid-, and high-frequency portions of the signal. In addition, the use of multiple denoising channels provides even greater flexibility and allows the wavelet denoising approach to extract rectangular signals over multiple decades of timescales. Operating on example signals from single Taq DNA polymerase molecules, the wavelet denoising procedure outlined here can flatten and reduce the RMS of the baseline more than any other method tested while still preserving the spikes and sharp transitions that characterize the enzymatic activity of the polymerase. Further analysis of denoised signals can include a state-finding procedure to determine the number of states, along with the frequency, durations, and other characteristics of the various states of the biomolecule.

CHAPTER 5

Characterizing Enzyme Motion with Features

5.1 Introduction

Single-molecule studies can observe the actions taken, and conformational states visited, by an individual biomolecule and to track the history of the biomolecule's activity over time. Characterizing the activity of a single biomolecule involves identifying the number of states and determining the distributions of the attributes of each state, including the distributions of durations and amplitudes of events in a particular state, as demonstrated in Figures 2.4, 3.8, 3.11, and Table 3.2. Though calculating distributions of state characteristics can help determine whether biomolecule activity is best modeled by Poisson statistics or other stochastic processes (38), there are significant advantages in accurately characterizing individual events. In particular, applications like DNA sequencing require accuracy both for counting of incorporated bases as well as the identity of each base (196, 197). In addition, biomolecules can exhibit multiple functions (89, 90, 107, 122), so accurate analysis of the biomolecule signal requires categorizing individual events by their corresponding function.

There are several challenges in using SWCNT-FET biosensors to accurately identify the various biomolecule conformations and categorize individual events.

First, enzymes exhibit both catalytic and non-catalytic conformational motions, and events corresponding to each may be distributed stochastically and difficult to separate. For

example, with T4 lysozyme, the $I(t)$ could be naturally separated into distinct segments in time, with each segment containing only catalytic or non-catalytic motion (198). In this case, counting the number of catalytic or non-catalytic events is a simple matter of counting the number of events in each segment. However, with other enzymes like the DNA polymerases, the catalytic and non-catalytic motions are intermixed in time. Figure 5.1 shows an example $I(t)$ from Taq DNA polymerase which contains both catalytic and non-catalytic events distributed stochastically in time. In this example, the non-catalytic events are shorter in duration than the catalytic events, by about an order of magnitude. However, the histograms in Figure 5.2 show that the distributions of catalytic and non-catalytic events overlap, so accurate categorization requires more information than only the duration of events.

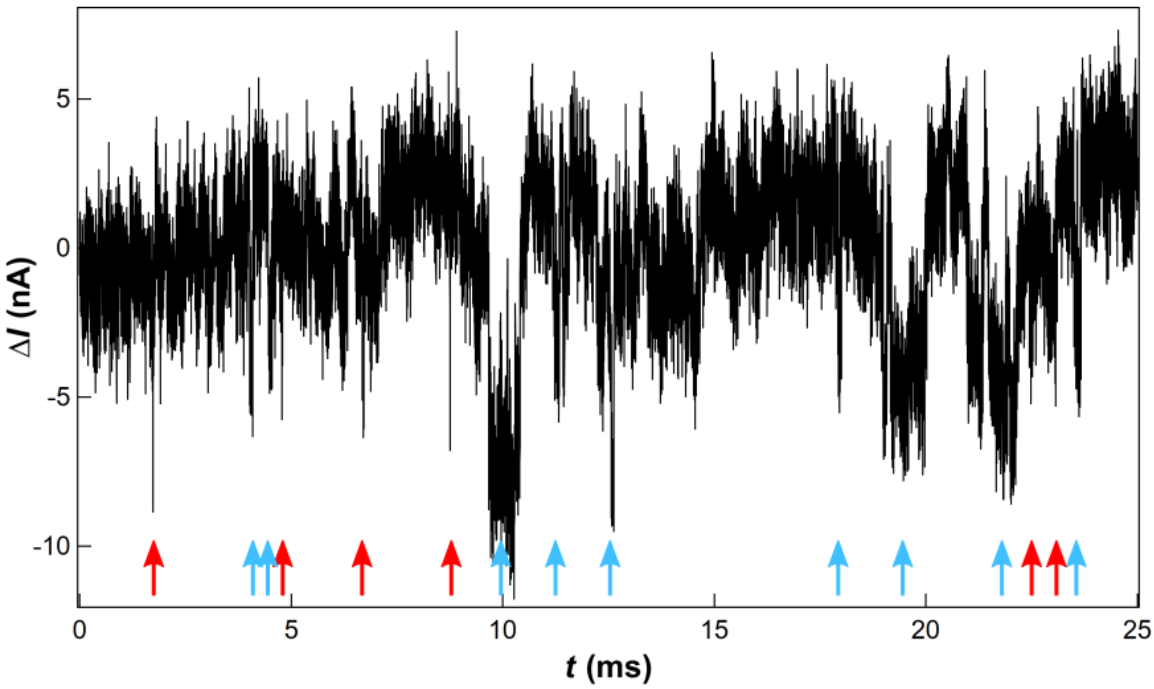

Figure 5.1: 25 ms increment of the raw $\Delta I(t)$, showing both short events (red arrows) and long events (light blue arrows) intermixed in time.

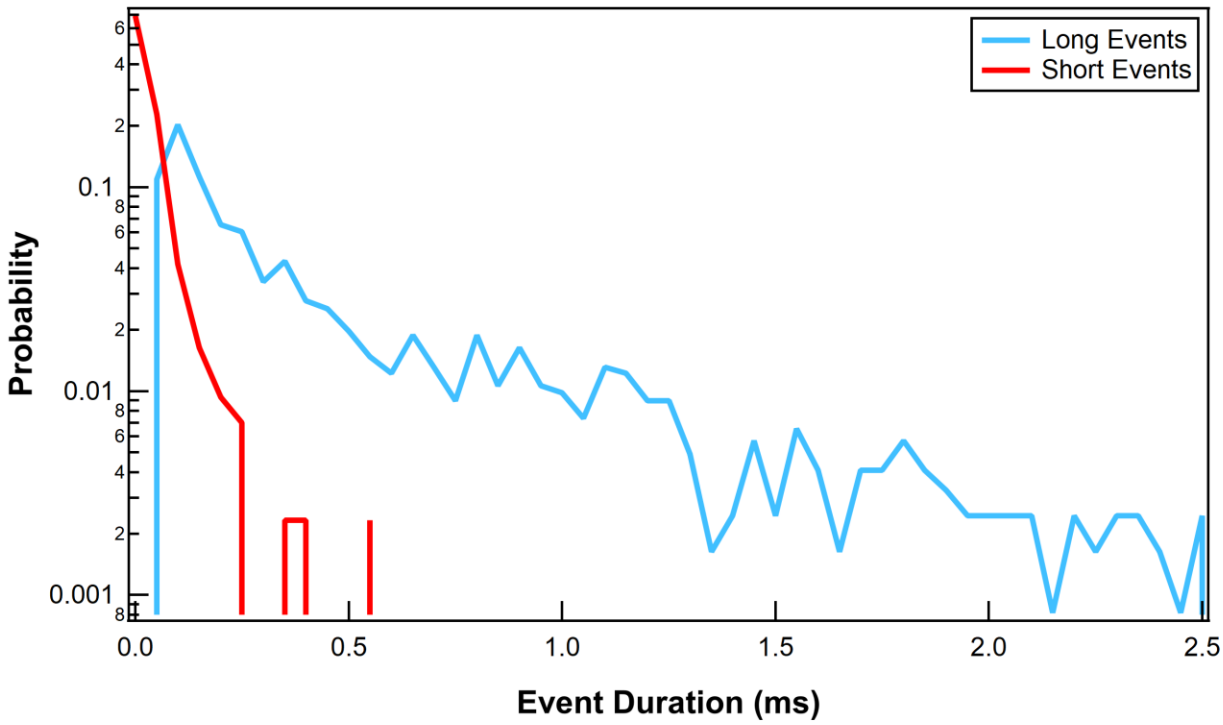

Figure 5.2: Probability distributions of the long (light blue) and short (red) events.

Second, multiple chemical processes can correspond to the same conformational motion of an enzyme, and multiple conformational motions may map to the same state in the signal. This is partly due to the lack of spatial information in the single-channel SWCNT-FET recording, which is analogous to an audio recording from a single microphone rather than from an array of microphones. Just like a single microphone may have trouble distinguishing between two identical speakers positioned at different spatial locations (142), a signal from a SWCNT-FET may not contain enough information to distinguish between two different processes solely based on the amplitude of an event. For instance, DNA polymerases use the same fingers-closing motion to complete catalysis of all four nucleotide bases, irrespective of the base identity (35, 102). Previous studies of the Klenow Fragment from DNA polymerase I showed that the events corresponding to each of these four bases exhibit

differing mean amplitudes and durations, but that the distributions of the amplitudes and durations overlap substantially (2). Given only this information, accurate identification of the identities of individual nucleotide bases is impossible. However, the distributions might separate with the addition of more details from the $I(t)$, especially if those details carry information about subtle differences in conformation or charge distribution.

Third, some events, henceforth called complex events, exhibit a shape or internal structure featuring sharp transitions or spikes within the duration of the event. An example of a complex event is shown in Figure 5.3A, contrasted with a simple event, which approximates a smooth rectangular wave, in Figure 5.3B. Since these mid-event spikes are similar to independent events, they can confuse event-identification and counting procedures, causing such procedures to split a single event into two or more separate events. Accurate counting of catalytic events requires that these events be counted only once, in their entirety, by the event-finding algorithm.

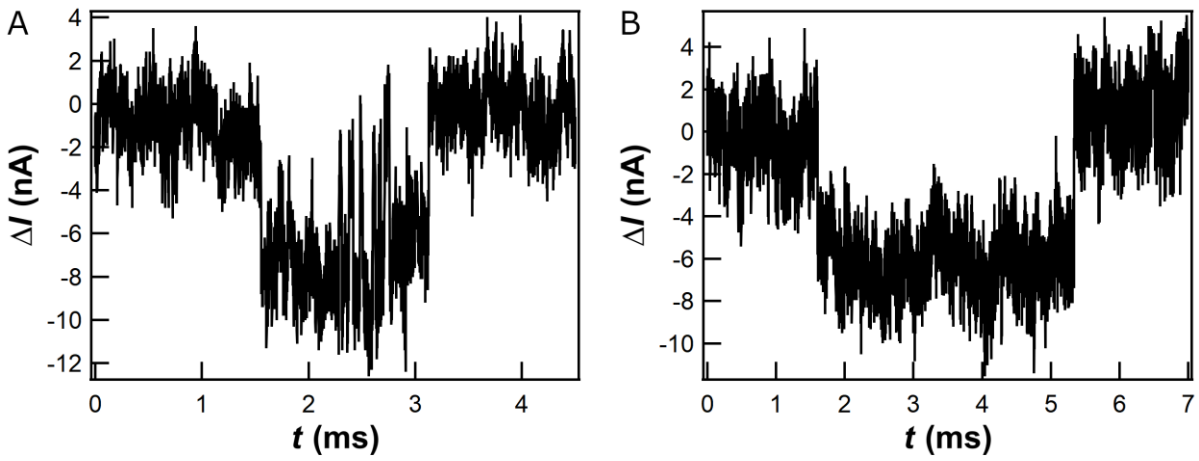

Figure 5.3: Examples of complex (A) and simple (B) long events.

Finally, identification of biologically-significant differences between two measurements is complicated by inherent variation in the measurements themselves. For instance, two measurements may contain signals from two different variants of the same enzyme, and the differences between the two measurements should ideally reflect the differences between the two enzyme variants. Unfortunately, $I(t)$ signals and events can vary between individual single molecules or sensors, and even separate measurement sessions of the same molecule, in ways that are independent of the biomolecule under observation. Individual biomolecules vary due to slight differences in temperature and conformation, and the characteristics of an individual molecule may vary from measurement to measurement due to dynamic disorder or degradation (39). Individual sensors may differ due to slight variations during fabrication (25). Individual measurements may be affected by different sources of noise. Thus, drawing accurate conclusions about the differences in signal between variants of a biomolecule, or a biomolecule's activity in the presence of various substrates, requires eliminating any differences that are non-biological in origin.

The task of identifying individual events and distinguishing between different functions requires more information than contained in 100 kHz $I(t)$ recordings at 25 kHz bandwidth (2), which was the experimental measurement limit of previous projects. To increase the information available for analysis, two approaches are being pursued simultaneously. First, on the experimental side, the effective measurement bandwidth has been increased to 1 MHz to capture as many details of the biomolecule dynamics as possible in the $I(t)$ recordings (26). Second, on the analysis side, the characteristics of each individual event are captured in a set of quantitative descriptors, which is stored as a state vector. The state vector, which

contains information such as event duration, amplitude, standard deviation of the $I(t)$ within an event, power in specific frequency bands, etc., can then be processed with linear algebra or machine learning techniques to look for correlations, separations of clusters, or other statistical relations between different events and states (199). Similar approaches have been successfully used to determine single-molecule kinetics (200-202) and to identify individual nucleotide bases (22, 196, 203-207) and amino acids (20, 208) in both DNA and protein sequencing, respectively.

This chapter describes the details of the second approach – the analysis and conversion of individual events to state vectors – in three parts. First, the procedure for identifying events is described, and definitions of the individual descriptors in the state vector, called features, are given. Second, an initial analysis of correlations and separation of clusters is presented, verified using two training sets. Third, the correlations and significant features are further characterized using principal component analysis, and pathways toward additional analysis with other machine learning techniques are described.

5.2 Event Identification and Feature Extraction

5.2.1 Automatic Event Identification

Identifying biomolecule events in the raw $I(t)$ involves two main processes: 1) denoising and detrending the $I(t)$ signal (discussed in Chapter 4), and 2) determining the locations and durations of spikes and transitions in the denoised signal. Chapter 4 describes how the signal

denoising makes the state positions consistent in time by flattening the baseline (removing low-frequency fluctuations) and reduces the noise amplitude within each state to clearly determine the boundaries between states. This section discusses the second process: determining the locations and durations of the spikes corresponding to biomolecule activity.

In the ideal case, identifying the position of each state and the boundaries between states are as simple as finding the number of significant peaks in a smoothed histogram of the denoised signal and finding the local minima between peaks. For instance, if the underlying signal contains two states, then the histogram of the denoised signal should show two peaks (Gaussian or otherwise), or at least a main peak (corresponding to the baseline) with a significant shoulder (corresponding to the second state). An example histogram of a 1 s increment of the wavelet low-pass denoised channel is shown in Figure 5.4A. The raw histogram (black) is smoothed (light blue) to round out the sharp peak at 0 nA which is a known feature of a wavelet-denoised signal (Section 4.4.4). The positions and widths of the peaks are calculated by fitting Gaussian functions (dark blue) to the peaks in the smoothed histogram, and the boundaries between states are defined by the intersection of the Gaussian functions in the valleys between states, as shown by the vertical red line in the middle. This method of identifying states works when there are enough sample points in the secondary states to reliably form a distinct peak that is separate from the main peak, which is true when each secondary state contains at least 2% of the sample points. The accuracy of this state identification method depends on the amount of fluctuation in the smoothed histogram and the accuracy of the curve-fitting procedure in identifying peaks.

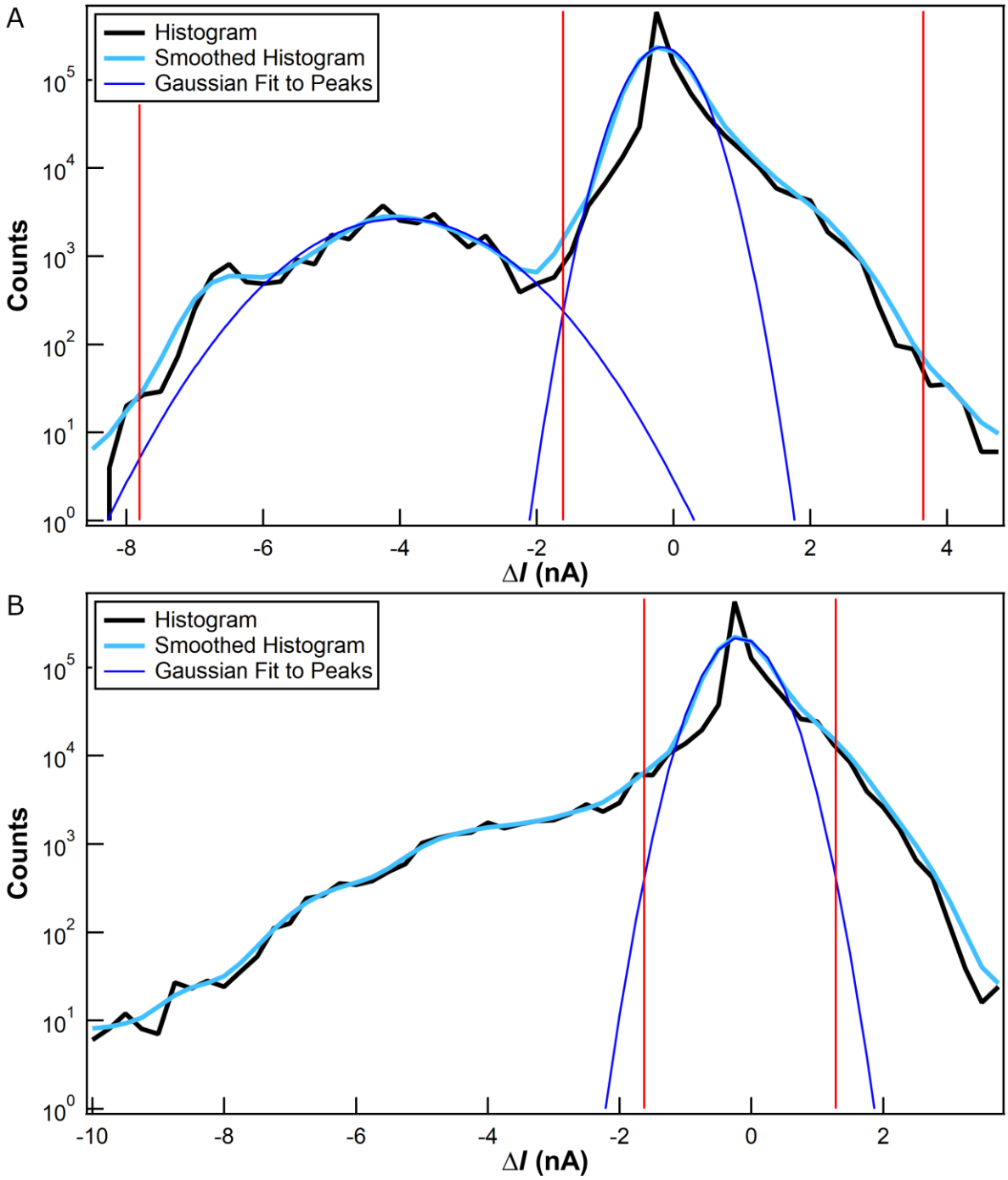

Figure 5.4: (A) Histogram (black) of a 1 s increment of the denoised $I(t)$, which shows two significant peaks. A smoothed version of the histogram (light blue) facilitates fitting the peaks to Gaussian functions (dark blue). (B) A similar histogram of another 1 s increment of the denoised $I(t)$, which shows only one significant peak.

When the $I(t)$ signal has a sparse distribution of events, there may be too few sample points in the secondary (non-baseline) states to form a distinct peak in the histogram, as shown in Figure 5.4B. This occurs when the secondary states contain less than 2% of the sample points. In this case, there is only one peak (corresponding to the baseline), and there is no valley to act as a clear boundary to mark secondary states. One way to identify the sparse events is to assume that there are always two "extrema" states on the outside edges of the histogram, as indicated in Figure 5.4A and B: a low state at the far left (lowest current position), and a high state at the far right (highest current position). In this way, the sparse events are still identified as residing in a secondary state, even if the histogram does not show peaks corresponding to additional states. The boundaries between these "extrema" states, marked by the vertical red lines on the left and right, can be set by defining a threshold, a distance from the baseline position, such that any samples beyond the threshold are assigned to the extrema state. The threshold is set as a multiple of either the median absolute deviation (MAD) of the denoised signal or the width of the fitted Gaussian function.

Once the number of states and the boundaries between the states have been identified, the state of each sample in the $I(t)$ is calculated, and consecutive samples residing in the same state are counted to calculate the duration of each state instance. If a state instance is shorter than some minimum duration, its samples are added to the previous state instance. Setting a minimum duration for an event is necessary to prevent counting of spikes that are too short to have been output from the electrical current amplifier. Such spikes might be due to electronic noise in the data acquisition card or artifacts in the denoising procedure, especially if wavelet denoising is utilized.

Proper characterization of complex events requires accurately identifying the start and end of the entire event, instead of identifying and characterizing only the individual spikes in the internal structure. A complex event can be considered as a combination of a cluster of individual events which are separated by short gaps, with the beginning defined by the start time of the first event and the ending defined by the end time of the last event. These individual events can be merged using a simple procedure: select two adjacent events, and if the waiting time between two events is less than 50% of the duration of either event, define a new event with the start time of the first and the end time of the second. This procedure can be repeated for every event in the entire measurement. By merging the individual short events into a single event, the complex event is counted as one, rather than multiple, events.

Finally, event identification must be customized for each denoising channel because each channel is designed to observe different characteristics of the signal. For instance, the wavelet high-pass denoising channel preserves events between 1 and 500 μs in duration, corresponding to the shorter, non-catalytic events, while the wavelet low-pass denoising channel preserves events between 50 μs and 10 ms in duration, which are predominantly catalytic. The minimum event duration for each denoising channel can be set to a value close to the minimum timescale preserved: 10 μs minimum for the high-pass channel and 80 μs minimum for the low-pass channel. In addition, the output of the low-pass channel has a lower baseline RMS value than the output of the high-pass channel, so the thresholds for calculating the extrema states will be lower for the high-pass channel because there is less noise to avoid. Finally, the low-pass channel is better suited for identifying complex events

because it smooths away some of the spikes and transitions exhibited by a complex event, so the analysis of the low-pass channel should also include the merging of adjacent events. By contrast, the high-pass channel is better suited for characterizing the individual short spikes and transitions within a complex event, so the merging of adjacent events should not be included in the analysis of the high-pass channel.

The procedure for automatically identifying events has the following steps:

1) Select a batch of denoised signal (usually 1 s of data, or 10^6 samples at 1 MHz)
2) Generate a histogram of the samples in that batch, then smooth the histogram
3) Determine the state positions and boundaries using the smoothed histogram
    a) Determine the peak centers and widths by fitting Gaussian functions to the peaks
    b) Calculate the boundaries for the extrema states:
        i. Multiply the width of the largest fitted Gaussian by the MAD multiplier to determine the threshold
        ii. Add the threshold to the position of the highest state to determine the high extrema boundary
        iii. Subtract the threshold from the position of the lowest state to determine the low extrema boundary
4) Assign each sample point to one of the states
5) Calculate the duration of each state occurrence
    a) Count the number of consecutive sample points belonging to the same state, for every occurrence of each state

b) If the duration of a state occurrence is lower than a minimum value, merge that state occurrence with the previous state occurrence

c) If the denoised signal is from a low-pass channel:

i. For each event (non-baseline state occurrence), look forward and backward in time by half of the event duration and search for events with the same state

ii. If events of the same state are found during the forward/backward search, merge those events with the current state, then repeat step 5c

6) Record the timestamp, state number, and duration of each state occurrence

5.2.2 Feature Extraction For SWCNT-FET Signals

Feature extraction is the process of converting raw measurement data into a collection of metrics, called features, that describe the important aspects of the data in a way that machine learning programs can understand (199). The values of the features need to be structured in a regular form that can be input into machine learning algorithms to look for correlations, clusters, or other patterns in the data. The typical form is a 2D matrix which is a collection of 1D feature vectors, each of which is composed of the values of the features for an individual measurement or object. Sometimes, the raw data possesses a consistent structure that allows it to be passed directly into a machine learning program, such as a dataset consisting of 1024 x 1024 pixel images of human faces for facial recognition (209). Feature extraction is crucial when the raw data is irregularly structured or unstructured (such as in

an audio recording of a speech in which the words are spaced irregularly in time), or when the dataset is too large to process in a reasonable time with available computing resources.

The types of useful features depend on the type and structure of the raw data. For instance, in 2D images of faces used for facial recognition, the color of the eyes, shape of the ears, and color of skin or hair are all features that can help identify an individual person (210). Similarly, in EEG and ECG sensing, the amplitude and shape of the individual spikes, as well as the power in specific frequencies and the presence of bumps before and after the spikes, are all characteristics of the signal which are used for medical diagnoses (211-213).

Initially, SWCNT-FET signals do not have any structure suitable for extracting features, since the signals consist of a single continuous sequence of $I(t)$ values with no intrinsic points for separations. Even when the $I(t)$ values within individual events corresponding to biomolecule activity are extracted, each event contains a different number of samples due to their differing durations, and thus lack the regular structure required for machine learning analysis. Thus, it is not possible to simply extract the $I(t)$ from each event to pass directly to further analysis routines. Instead, each event must be characterized by features that output a consistent number of values regardless of the number of samples in the original event.

Though the SWCNT-FET signals are in the time domain, features can be extracted both from the time domain and the frequency or wavelet domain. In fact, it is useful to extract two sets of features from each domain: one corresponding to the raw signal, and one corresponding to the denoised signal. Since the spikes corresponding to biomolecule activity are localized

in time, the wavelet transform (particularly the UWT) is better suited to capture characteristics of the spike than the Fourier transform, since the latter is too delocalized to be sensitive to a single spike. For typical 1 MHz $I(t)$ recordings, the UWT features were extracted from scales 2-10, which correspond approximately to frequencies 250 kHz and 1 kHz, respectively. Most of the features focus on characterizing the data within an event, excluding the first 10% and last 10% of the duration of an event to avoid including the transition between events. However, the transitions themselves contain useful information that can be captured with additional features. A description of the types of features extracted is provided in Table 5.1. A full list of features is provided in Table A.1 in Appendix A.

Feature Name	**Feature Description**
Event Timing	
Event Duration	Duration of event
Raw $I(t)$ / Denoised $I(t)$ within an event	
Max	Maximum value of the $I(t)$
Min	Minimum value of the $I(t)$
Mean	Mean of the $I(t)$ values
Standard Deviation	Standard deviation of the $I(t)$ values
Amplitude by Mean	Difference between the mean of the current event and the previous event
Median	Median of the $I(t)$ values
Lower Quartile	Value at the lower quartile of the values in the $I(t)$
Upper Quartile	Value at the upper quartile of the values in the $I(t)$
Median Absolute Deviation (MAD)	Median absolute variation of the $I(t)$ values
Amplitude by Median	Difference between the median of the current event and the previous event
Total Variation	Total variation, or distance traveled, by the $I(t)$ values
Total Variation / Time	$I(t)$ Total Variation feature divided by Event Duration feature
Skewness	Skew of the $I(t)$ values
Kurtosis	Kurtosis of the $I(t)$ values
Square Root of Sum of FFT Power	Square root of the sum of the amplitudes of the PSD of the $I(t)$ values within an event

UWT Coefficients (scales 2-10) within an event	
Mean	Mean of the UWT coefficients
Standard Deviation	Standard deviation of the UWT coefficients
Median	Median of the UWT coefficients
Median Absolute Deviation (MAD)	Median absolute deviation of the UWT coefficients
Total Variation	Total variation, or amount of distance traveled, by the UWT coefficients
Total Variation / Time	UWT Total Variation feature divided by Event Duration feature
Skewness	Skewness of the UWT coefficients
Kurtosis	Kurtosis of the UWT coefficients
Number of Zeros	Number of times the coefficients in the UWT scale level crosses zero
Number of Zeros / Time	Number of Zeros feature divided by Event Duration feature
Amplitude of Top 3 Peaks	Value of the peak height for the top three peaks in the UWT within a single event
Amplitude of Top 3 Valleys	Value of the valley height for the top three valleys in the UWT within a single event
Features From Multiple Channels	
Number of High-Pass Events Within a Low-Pass Event	Number of events detected by the high-pass channel within the duration of an event detected by the low-pass channel

Table 5.1: Description of features used to characterize individual events.

5.3 Correlations Among Event Features

Many of the selected features are proportional to event duration. These include: the total variation of the raw $I(t)$, the total variation of the raw UWT scale levels, and the number of times the plot of the raw UWT coefficients crosses zero. Figure 5.5A plots two features relative to event duration: the total variation of the raw $I(t)$ within an event (black dots), and the total variation of the UWT coefficients, in scale 5, within the same events (red dots). The data are plotted on a log-log scale. For both features, the individual points are distributed roughly linearly with a slope of 1, indicating a linear relation with event duration.

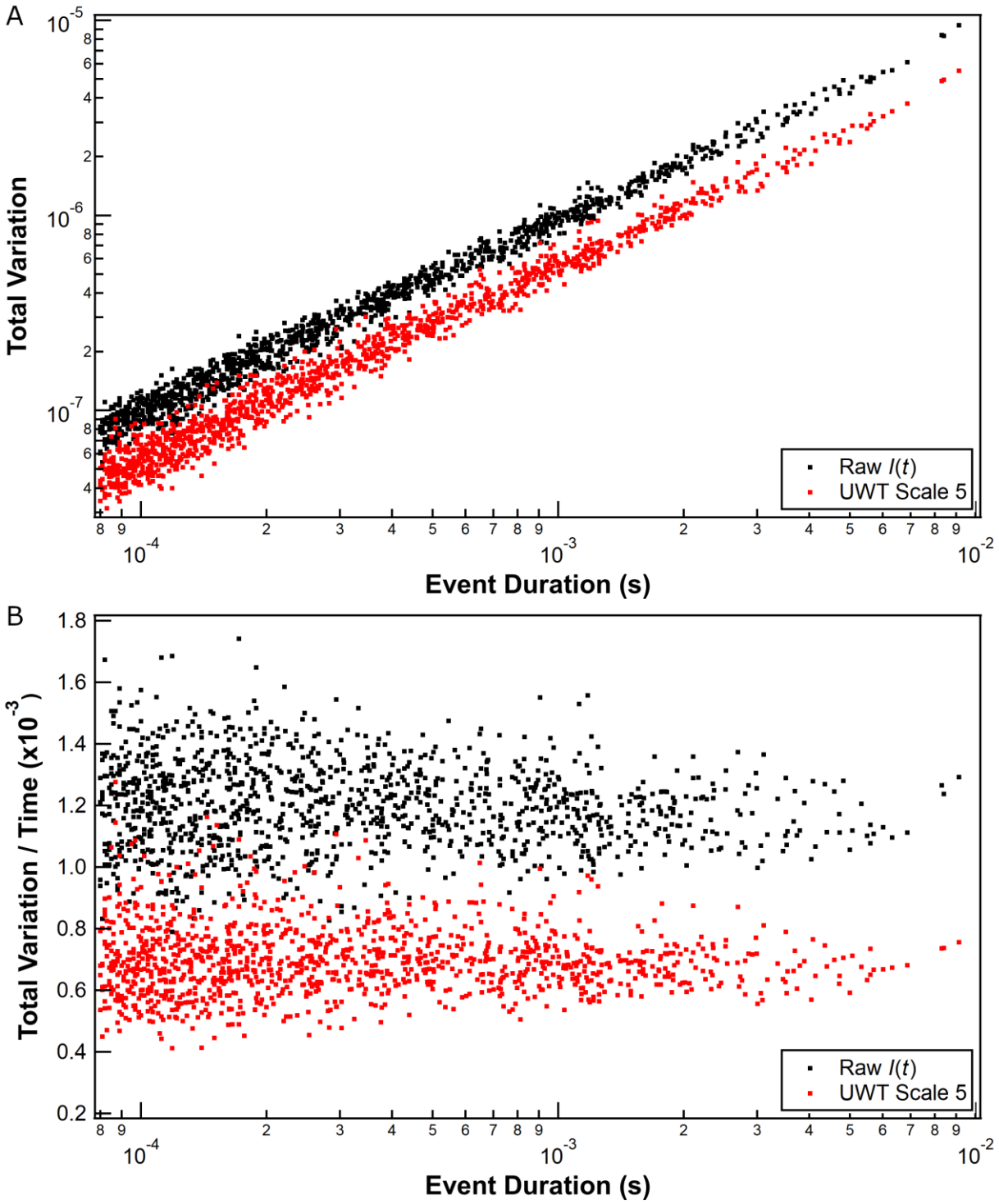

Figure 5.5: (A) Plots of two features compared to event duration. The two features are: total variation of the raw $I(t)$ within an event, and total variation of the UWT coefficients, in scale 5, within the same events. Both features are proportional to event duration, as shown by the slope of 1. (B) For both features, an additional feature is defined by dividing the feature by the event duration, creating features uncorrelated with event duration.

For each of the features proportional with event duration, an additional feature is defined by dividing the feature value by event duration, decorrelating this new feature from event duration. For example, Figure 5.5B shows the same data as in Figure 5.5A but with each value divided by the duration of the corresponding event. The resulting feature is uncorrelated with event duration.

A linear relation with event duration is primarily due to the presence of noise in the signal. Figure 5.6A displays a segment of raw $I(t)$, along with the same signal processed with the two wavelet denoising procedures defined in Chapter 4: the wavelet high-pass channel (displayed in light blue) and the wavelet low-pass channel (displayed in red). The plot of the $I(t)$ of the low-pass channel is smooth and lacks high-frequency oscillations, while the other two $I(t)$ signals retain some high-frequency noise. Figure 5.6B plots the total variation of the $I(t)$ within an event (for the raw, high-pass denoised, and low-pass denoised channels) against the duration of the event, comparing the features of the same events when processed using the two denoising channels. The distribution resulting from the high-pass denoising is linear with respect to event duration (slope of 1 on a log-log plot), matching the relation exhibited by the total variation of the raw $I(t)$.

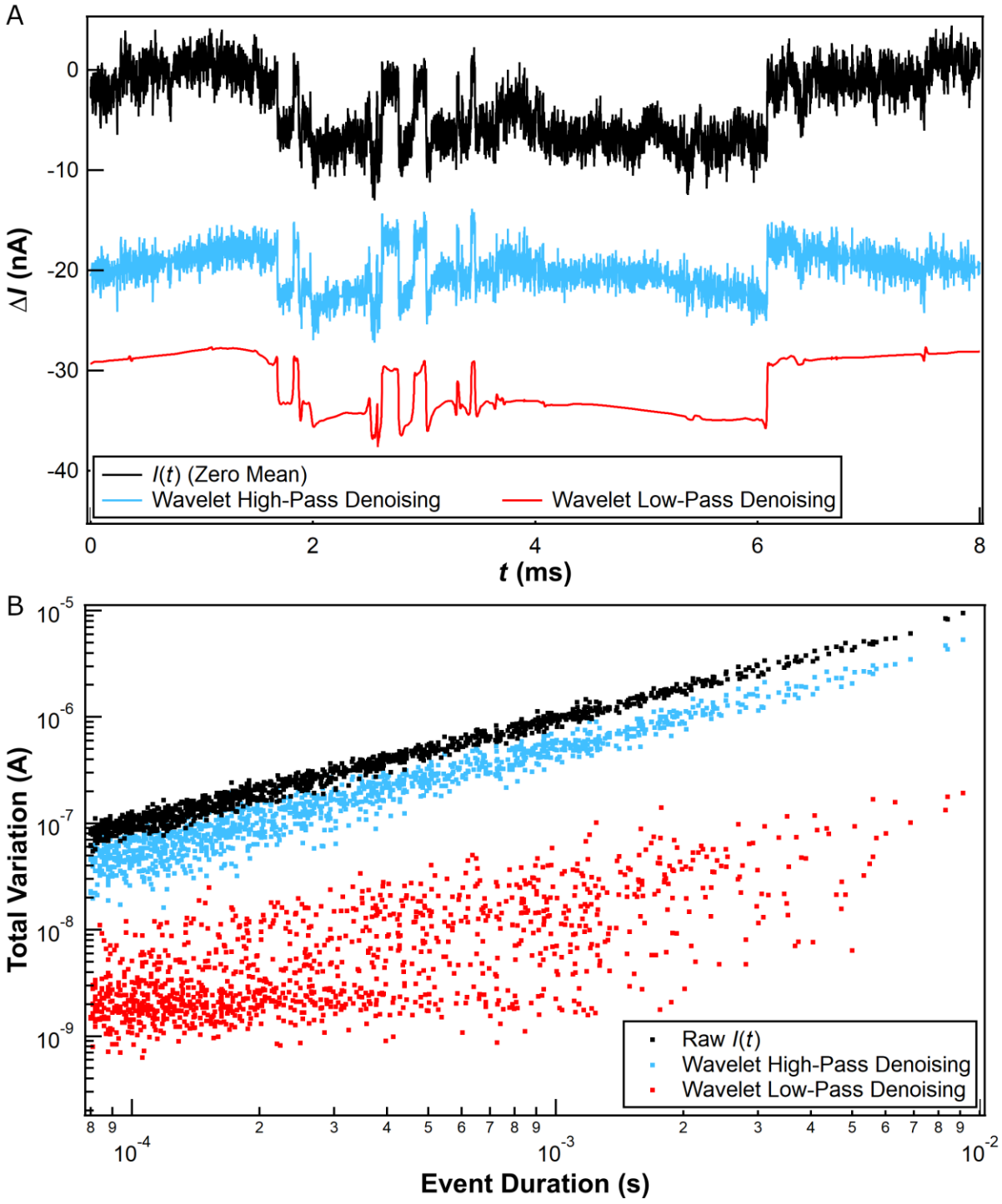

Figure 5.6: (A) Examples of $I(t)$ from the wavelet high-pass (light blue) and low-pass (red) denoising, offset from the raw $I(t)$ signal (black) by the tick values, as originally shown in Figure 4.21. (B) Comparison of the total variation of the raw (black), high-pass (light blue) and low-pass (red) denoised $I(t)$ signals.

By contrast, the distribution for the low-pass denoising possesses a weaker, non-linear relation with event duration. The distribution for the low-pass denoising exhibits lower total variation than that for the high-pass denoising (about an order of magnitude lower), which is expected due to the more aggressive smoothing in the low-pass channel. The low-pass denoised $I(t)$ lacks high-frequency noise and is not linearly correlated with event duration, so the linear relation is likely due to the presence of high-frequency noise, which contributes to the total variation at a constant rate. The distribution of low-pass denoised features also exhibits greater spread, suggesting that the total variation of the low-pass denoised signal contains additional information that is hidden by noise in other channels.

Some features exhibit little correlation with any other features. This may be because the features contain no useful information or because the useful information is hidden by noise or signal processing artifacts. Figure 5.7 shows a comparison between two such features.

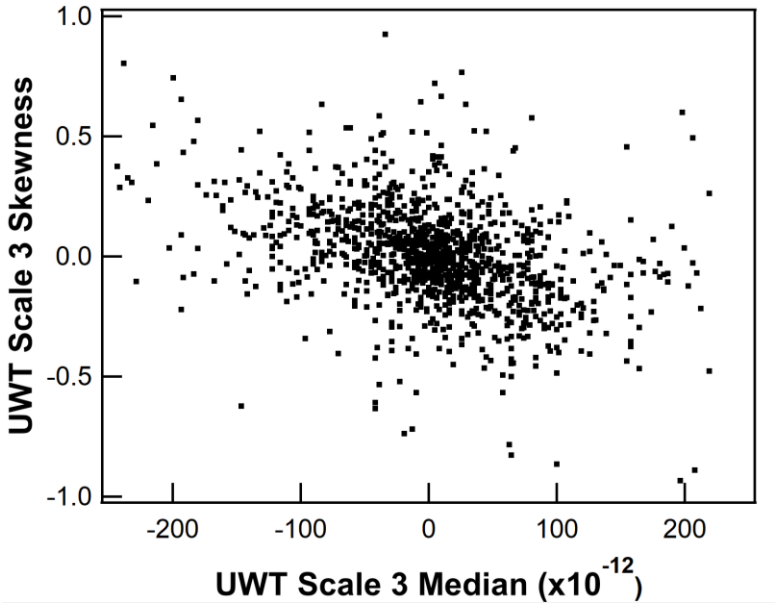

Figure 5.7: Features that show no correlation with any other features.

The correlations between individual features is illustrated as a correlation matrix in Figure 5.8, with positive correlation (+0.1 to +1.0) in dark red to orange, negative correlation (-1.0 to -0.1) in green and blue, and no correlation (-0.1 to +0.1) in yellow. The features are organized in blocks, with the size and general content of each block indicated by the lines and labels at the top of the figure. Many of the positive and negative correlations are connected to event duration (feature number 1). The large red triangle at the bottom right reflects the correlations in the amplitudes of the largest peaks in the UWT scale levels within the duration of a single event. The yellow and red pattern of squares in the middle reflects the correlations between UWT scale levels of the mean, standard deviation, median, MAD, skewness, kurtosis, total variation, and number of zero crossings. The dark red squares in the UWT total variation and UWT number of zero crossings blocks reflect the large (> 0.9) linear correlations between these features and event duration.

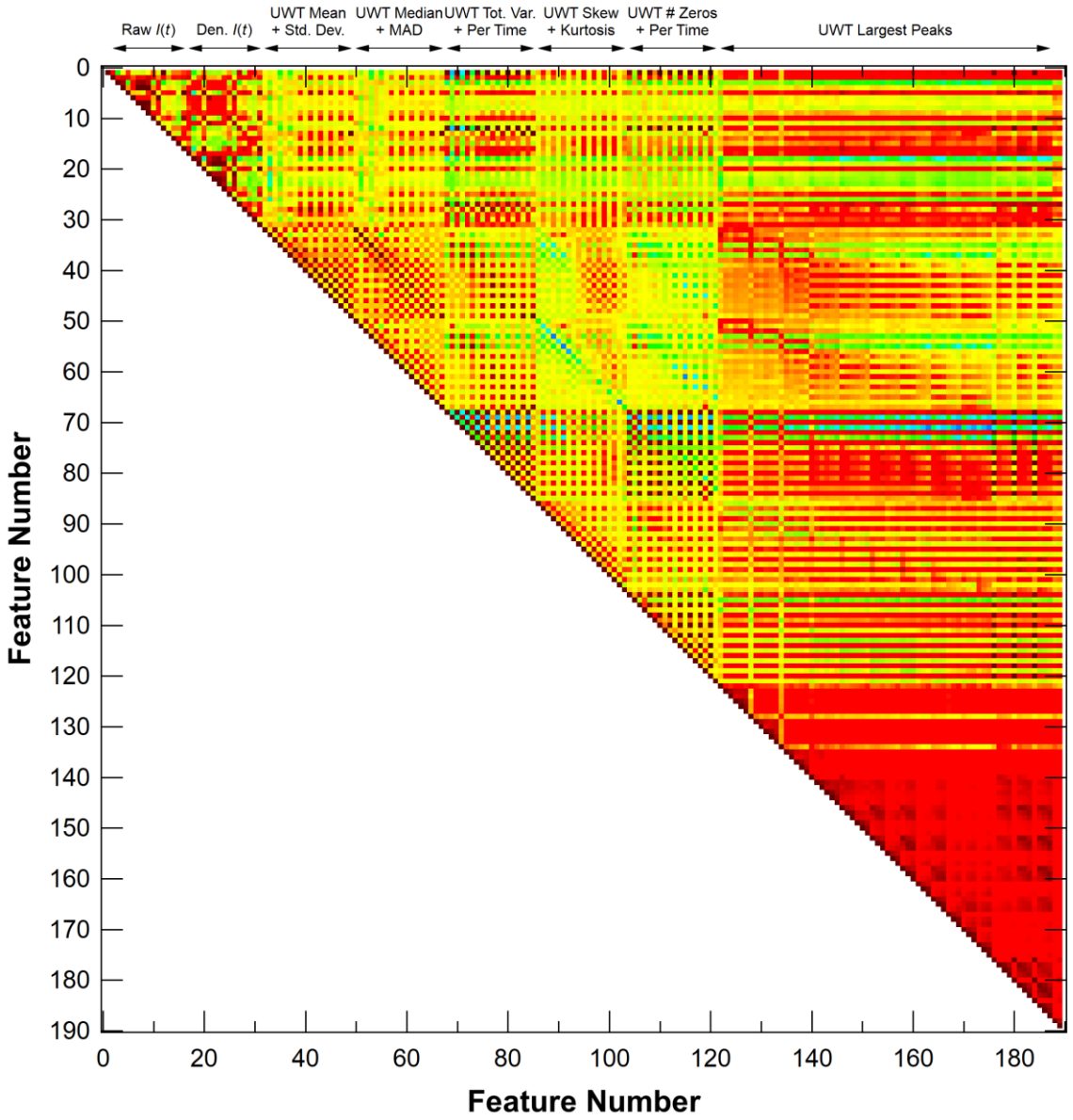

Figure 5.8: Correlation matrix showing the correlations between all 189 features. Positive correlations are shown in dark red and red and orange, negative correlations in green and blue, and no correlations in yellow.

5.4 Separation and Clustering in Event Features

A simple approach to quantifying the differences between events in separate measurements is to compare the distributions of the features from each measurement. The distributions

can be compared using standard statistical measures (such as the mean and standard deviation) or by displaying the individual features on a scatter plot. Significant differences will appear as a separation of peaks (when comparing distributions of only one feature) or as a separation of clusters (when comparing distributions for two or more features).

The clearest separations in features occur when there are systematic differences in measurement conditions between datasets. These systematic differences might reflect significant changes in biomolecule activity (such as a higher enzymatic rate due to higher temperatures), but also could be due to changes unrelated to biomolecule activity, such as changes in the measurement noise background. Typically, systematic conditions that affect the entire measurement can be characterized using multiple methods, so cross-checking by looking for similar patterns in related metrics can help determine the reason for the separations. Since some feature separations may be due to factors that are unrelated to actual biomolecule activity, each separation needs to be carefully investigated to determine the reason for the separation. Otherwise, naïve applications of clustering, support vector machines, or other machine learning techniques might falsely suggest that a feature shows biologically-significant separations when it does not.

The following discussion compares the events from two measurements of two different variants of Taq DNA polymerase, using two different sensors and measured on different days. The first measurement is of Taq mutant R695C, acquired using device PCB48-KJ 11bc on 3/4/2019 (timestamp 15.42.07), and is labeled PCB48-KJ in subsequent discussion. The second measurement is of Taq mutant R411C, acquired using device PCB48-JK 10cd on

3/14/2019 (timestamp 16.23.49), and is labeled PCB48-JK. The only difference between the two mutants is the position on the enzyme where it is anchored to the SWCNT-FET.

For example, if the high-frequency electrical noise is different between two measurements, there will be differences in the values of the total variation per unit time and the number of zero crossings per unit time for the high-frequency UWT scales. Figure 5.9 compares two noise-related features for scales 2 and 3 (corresponding to ~250 and ~125 kHz), with the total variation per unit time on the left axis and the number of zero crossings of the UWT scale per unit time on the bottom axis. The events from PCB48-KJ are shown in black, while events from PCB48-JK are shown in red. The plot with UWT scale 3 is on top (Figure 5.9A), while the plot with UWT scale 2 is on the bottom (Figure 5.9B). Figure 5.9A shows that, for scale 2, the red cluster is offset vertically and to the right from the black cluster, such that the best dividing line between the two clusters is a diagonal line slanting downward and slightly right. Figure 5.9B shows that, for scale 3, the black and red clusters are generally intermixed with similar vertical and horizontal centers. Similar comparisons made using other scales produce plots like Figure 5.9B. This implies that the measurement of PCB48-JK contains higher average power in scale 2 than the measurement of PCB48-KJ while containing similar power in the remaining scales.

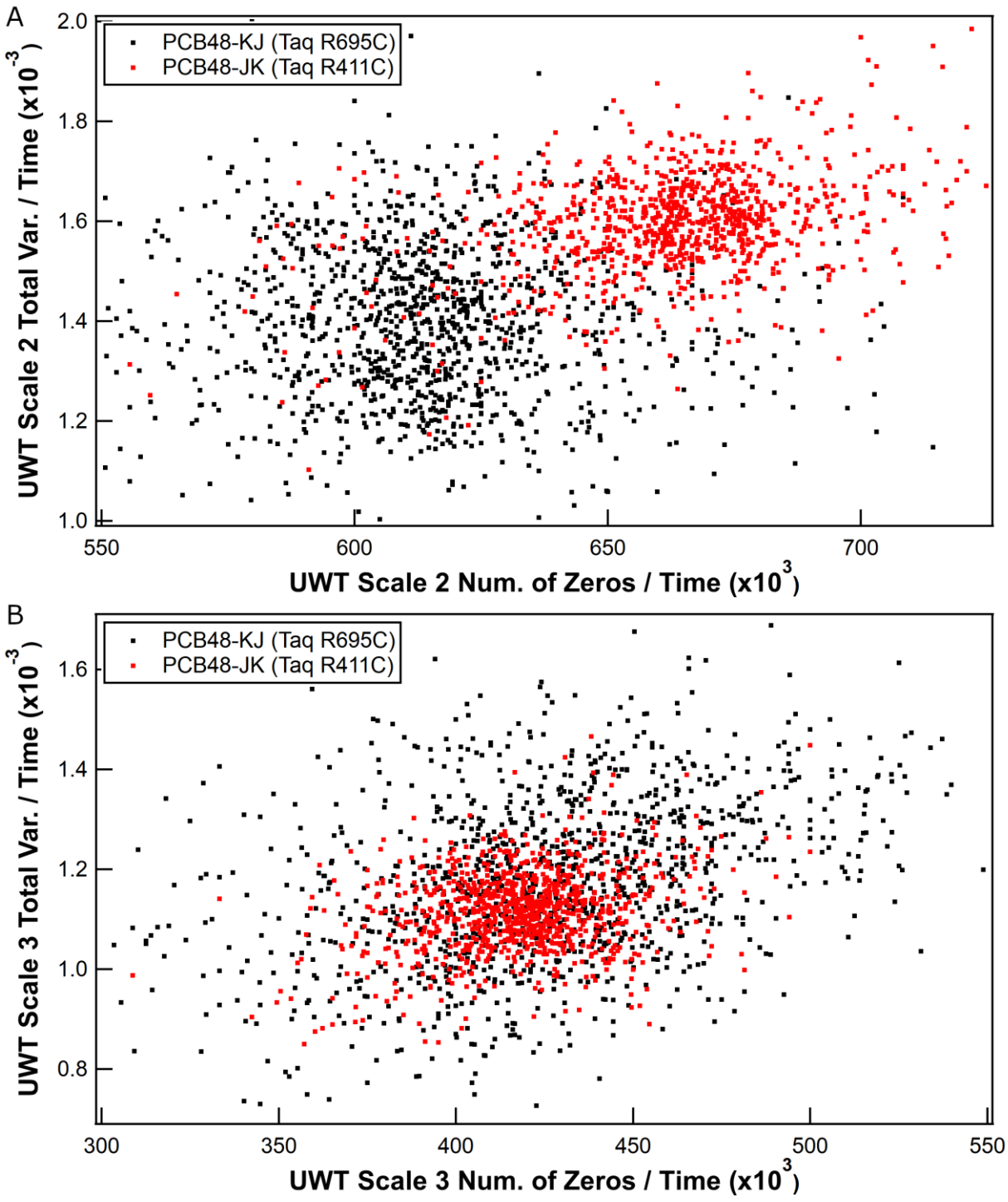

Figure 5.9: Scatter plots comparing two features (total variation per unit time and number of zeros per unit time) for events corresponding to PCB48-KJ and PCB48-JK for two UWT scale levels: (A) 2 and (B) 3.

Figure 5.10 shows four plots displaying histograms of the two features of Figure 5.9 for the same two scales (2 and 3) and the same two measurements (PCB48-KJ and PCB48-JK). In the plots corresponding to scale 3, the histograms from the two datasets exhibit significant (>80%) overlap, indicating that the power of the associated frequencies are similar. The histograms of the features for higher scales (not pictured) exhibit a similar overlap. However, the plots for scale 2 show a significant difference (<30% overlap) in the distribution of the features from the two measurements, especially for the number of zero crossings feature. For both features in scale 2, the peak for PCB48-JK is shifted to the right compared to the peak for PCB48-KJ, indicating that the power of the corresponding frequencies is higher in the first compared to the second.

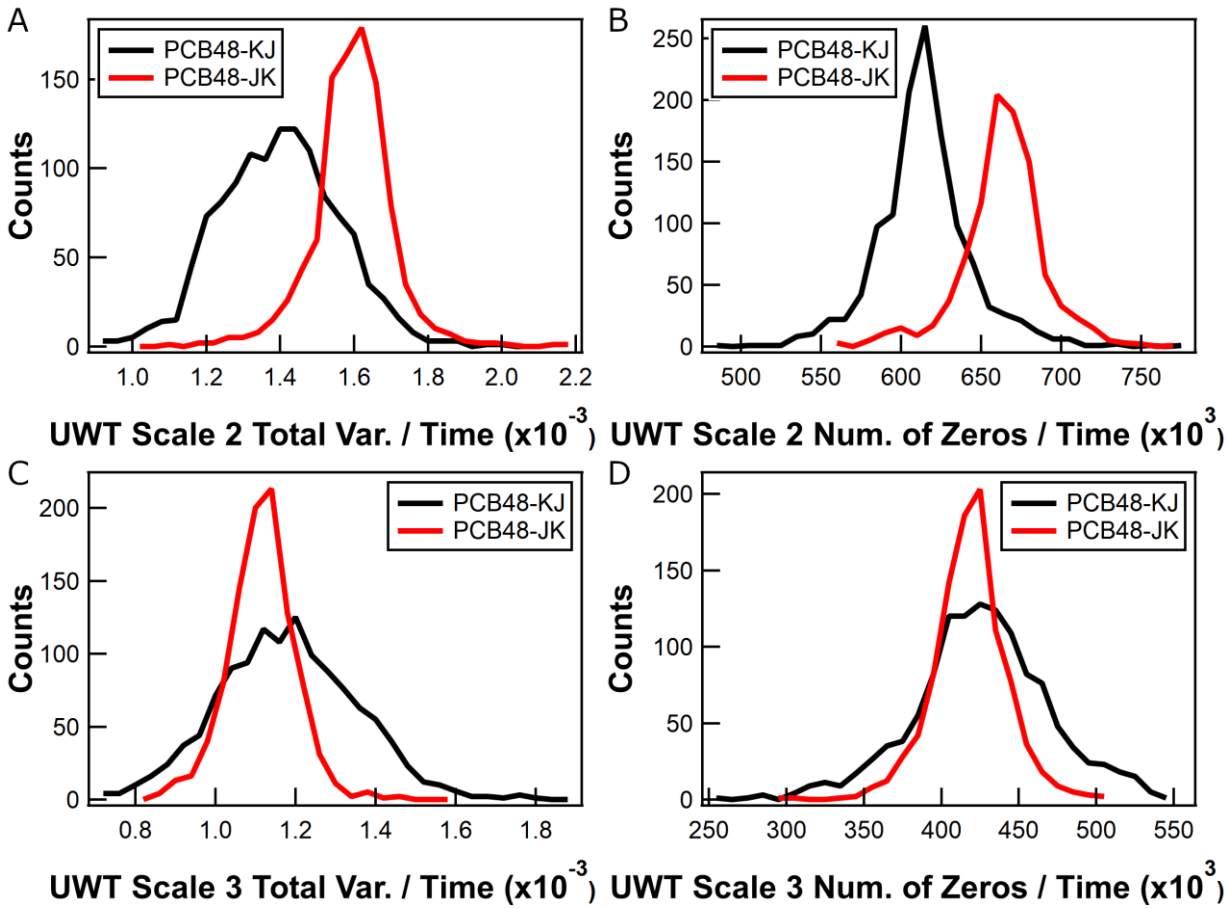

Figure 5.10: Histograms of the event features from PCB48-KJ and PCB48-JK: total variation per unit time for UWT scales 2 (A) and 3 (B), and the number of zeros per unit time for UWT scales 2 (C) and 3 (D).

The increased high-frequency power in PCB48-JK is easy to verify by looking at the power spectral density (PSD) plots from the two measurements, which is shown on a log-log scale in Figure 5.11. PCB48-KJ is shown in black and PCB48-JK (which is offset by a factor of 100) in red. Scale 2 corresponds to a frequency of ~250 kHz, while scale 3 corresponds to a frequency of ~125 kHz. At frequencies of ~260 kHz, ~350 kHz, ~400 kHz, and ~470 kHz, the PSD of PCB48-JK exhibits clusters of larger-amplitude spikes, indicating higher power at those frequencies, while the PSD of PCB48-KJ shows discrete spikes at intervals of 8 kHz instead of large spike clusters. The spike clusters in PCB48-JK were later experimentally

determined to be caused by high-frequency noise originating from the electronics for heating the buffer solution surrounding the SWCNT-FET sensor, not due to biomolecule activity. Thus, even though the two measurements acquired signals from different variants of Taq DNA polymerase and show differences in the distributions of features, some of those differences are not correlated with the variant of Taq under observation.

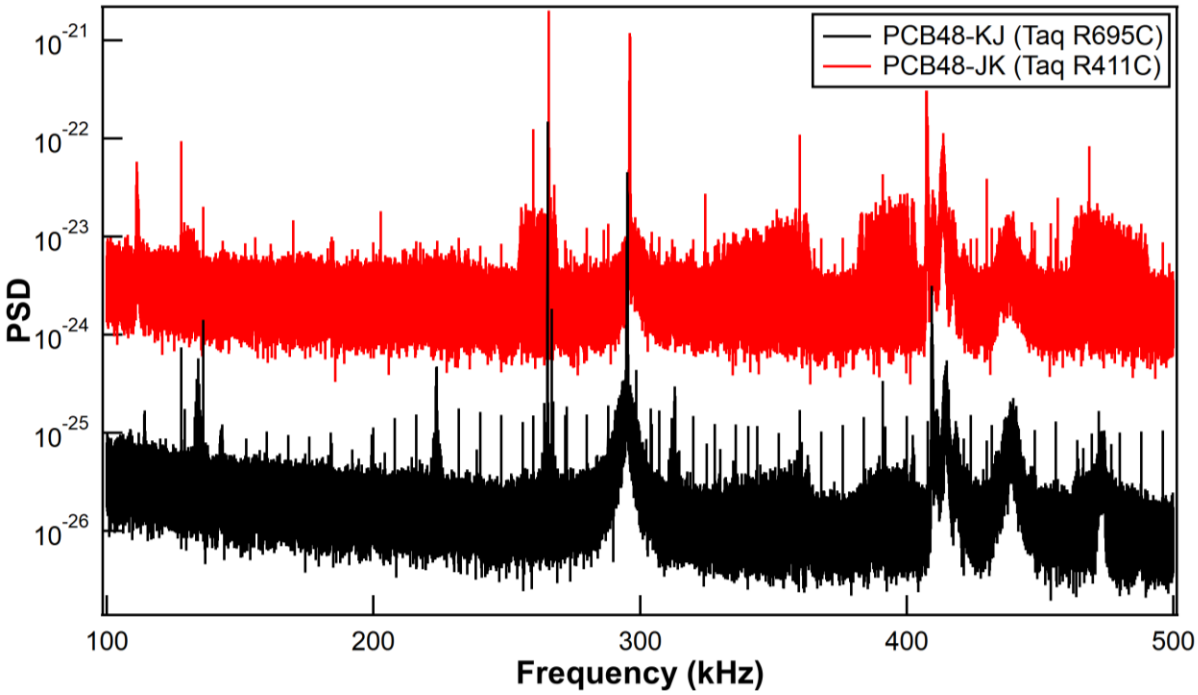

Figure 5.11: PSD of the raw $I(t)$ from PCB48-KJ and PCB48-JK.

Another example of a systematic difference in features between datasets is related to the amplitude of the event, which measures the distance of the second state from the baseline state. Figure 5.12A shows a histogram of the event amplitude (as measured from the baseline state using the median of each state, using the denoised signal) from the same two measurements previously mentioned. The histogram corresponding to the amplitudes of PCB48-JK has a peak at ~-3 nA, while the histogram for PCB48-KJ has a peak at ~-4.5 nA.

Since the average amplitude of events from PCB48-KJ is larger than the average amplitude from PCB48-JK, there should also be a noticeable difference in the $\Delta I(t)$ distribution. Figure 5.12B shows histograms of 2s segments of $\Delta I(t)$ from each dataset. Each histogram contains at least two peaks: a sharp primary peak near 0 nA corresponding to the baseline state, and a rounder, smaller-amplitude secondary peak (corresponding to the second state) to the left of the primary peak. The histogram of PCB48-KJ has potentially three peaks (primary at 0 nA, secondary at ~-4 nA, tertiary at ~-6.5 nA). The secondary peak of PCB48-JK (at ~-2.5 nA) is less negative and closer to the baseline than that of PCB48-KJ, which follows the same trend as the peaks in the distributions of event amplitudes.

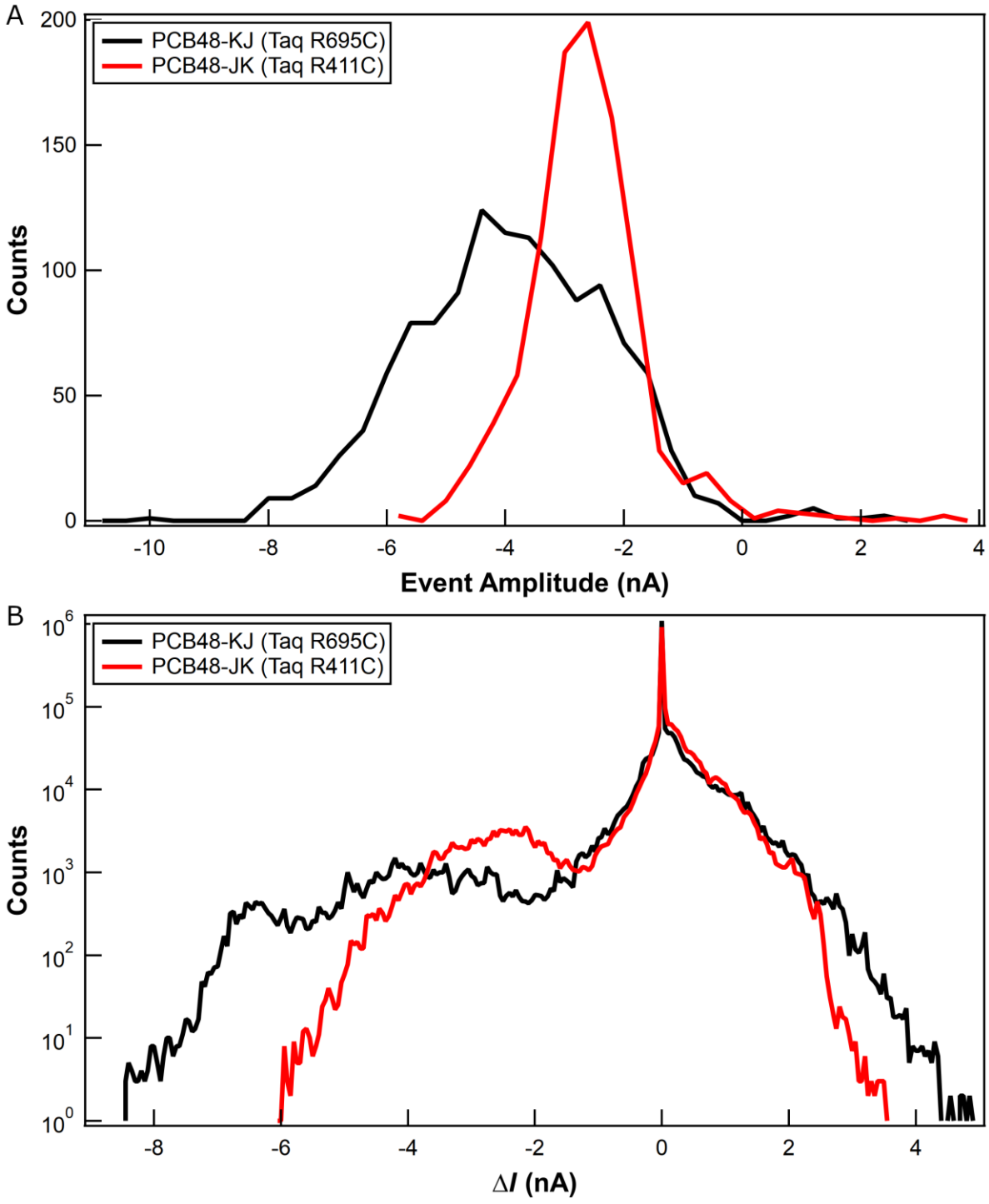

Figure 5.12: (A) Histograms of the event amplitudes (defined by the difference between the medians of the baseline state and second state) for PCB48-KJ and PCB48-JK. (B) Histograms of 2 s increments of the denoised $I(t)$ from PCB48-KJ and PCB48-JK.

The event amplitudes depend on the transconductance $\left(G = \frac{\Delta V_g}{\Delta I}\right)$ of the SWCNT-FET sensor, which varies from one sensor to another. Thus, proper comparisons of event amplitudes between measurements from different sensors requires converting the amplitudes to another metric that is robust to environmental changes and unaffected by sensor characteristics. Previous studies of SWCNT-FET sensor measurements of T4 lysozyme showed that the change in effective gate potential, $\Delta V_g = \Delta I * \frac{\Delta V_g}{\Delta I} = \Delta I * G$, remained consistent across different sensors and individual lysozyme molecules. When the histograms of both the event amplitudes (Figure 5.13A) and $\Delta I(t)$ (Figure 5.13B) are converted to ΔV_g, the differences between the two measurements increase by a factor of ~2.5. The event ΔV_g amplitudes peak at ~-50 mV for PCB48-KJ and ~-13 mV for PCB48-JK (Figure 5.13A), while the ΔV_g from the baseline peaks at ~-40 mV and ~-70 mV for PCB48-KJ and ~-13 mV for PCB48-JK (Figure 5.13B).

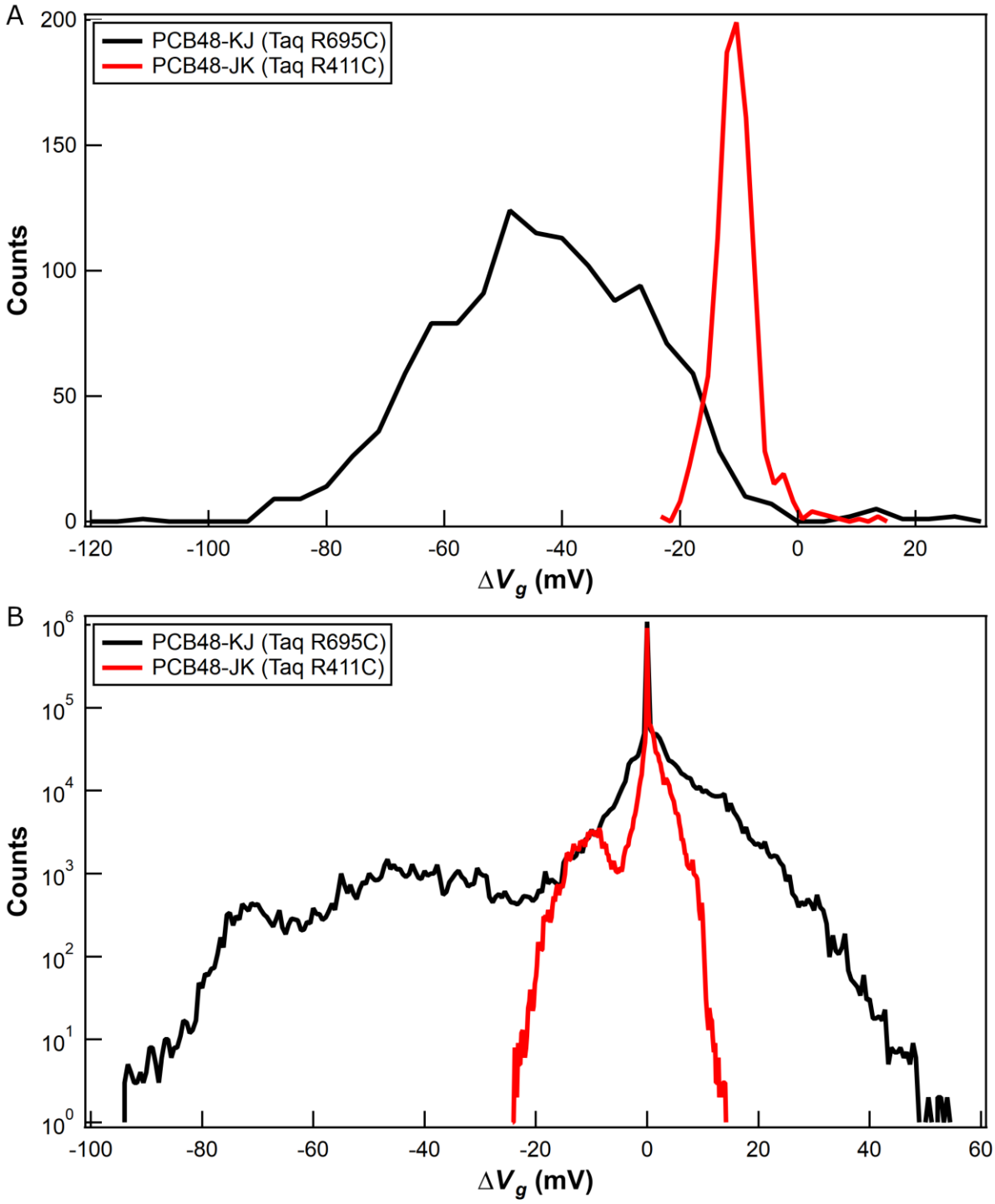

Figure 5.13: (A) Histograms of the ΔV_g calculated from the event amplitudes (defined by the difference between the medians of the baseline state and second state) for PCB48-KJ and PCB48-JK. (B) Histograms of the ΔV_g calculated from 2 s increments of the denoised $I(t)$ from PCB48-KJ and PCB48-JK.

The differences in ΔV_g between these two measurements, unaffected by the variations between individual SWCNT-FET sensors, show that the Taq R695C mutant produces events with amplitude ~2.5x greater than the Taq R411C mutant. Previous work with T4 lysozyme showed that the change in gate potential is directly proportional to both the mechanical displacement and charge magnitude of charged residues close to the SWCNT-FET during a change in biomolecule conformation. Thus, the differences in the ΔV_g features reflect a real difference in the motion of charges at the two attachment points. Further analysis of the structure of Taq DNA polymerase at these two locations may establish which charged residues are responsible for generating the events in the SWCNT-FET signal.

5.5 Characteristics of Event Features for Taq DNA Polymerase

When measuring the activity of Taq DNA polymerase (at the R695C attachment point) on a homopolymer template (polyT42) and its complementary nucleotide (dATP), three different categories of events are observed (Figure 5.14): short events (duration < 80 μs), simple long events (duration ≥ 80 μs), and complex long events (duration ≥ 80 μs with some substructure). Both short events and simple long events are generally rectangular-shaped pulses, with little current variation in the second state and sharp transitions (< 20 μs) between states. The complex long events are composed of a series of stepwise transitions or generally-rectangular pulses, sometimes separated by short gaps (< 80 μs) in which the signal completely returns to the baseline state.

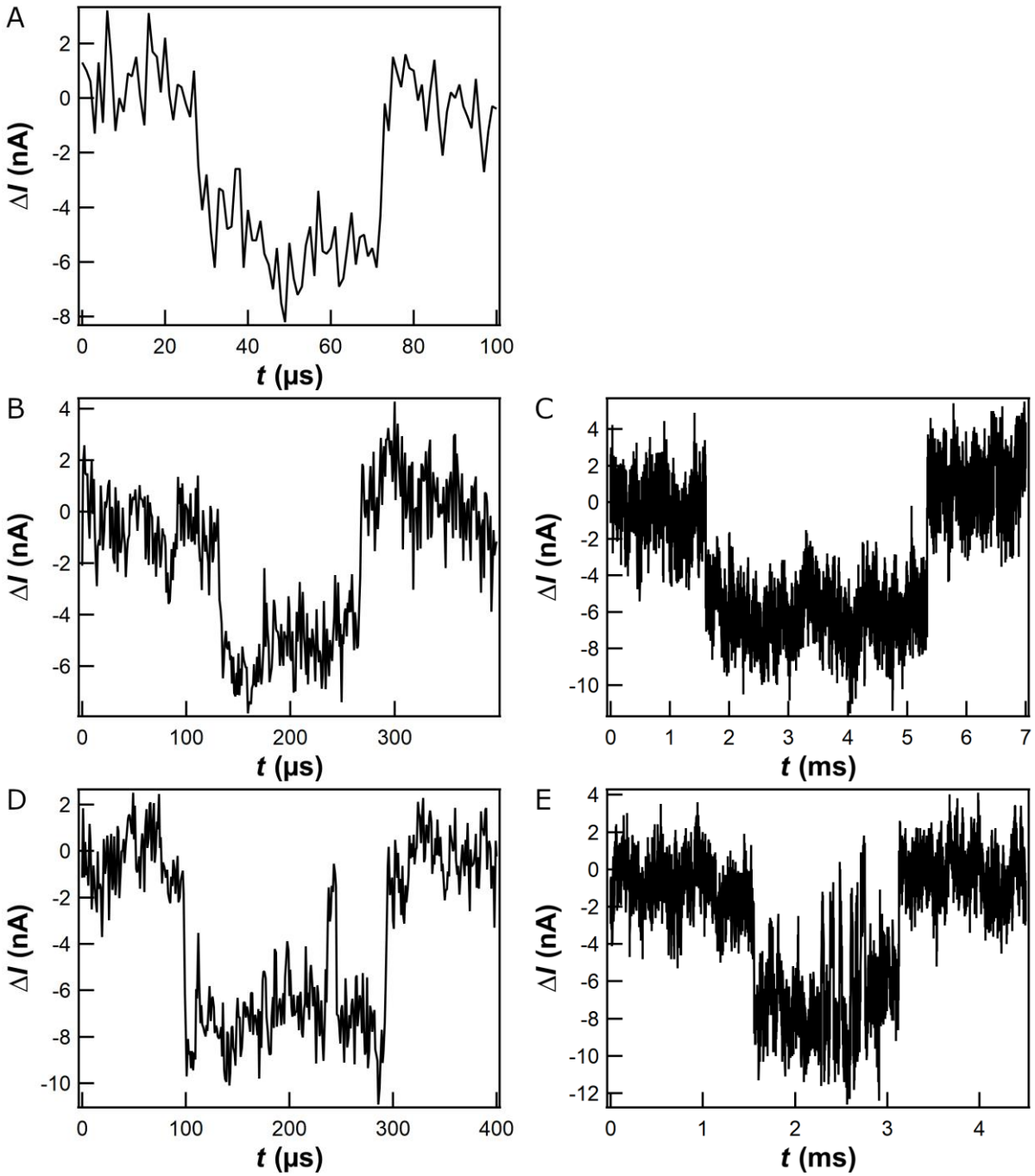

Figure 5.14: Short (A) and long events which are simple (B,C) and complex (D,E).

Previous experiments determined that the short events are not directly correlated with nucleotide incorporations nor with the rejection of non-complementary nucleotides but were otherwise unable to determine the cause of the short events. By contrast, the long

events were correlated with nucleotide incorporations, but the reasons for the substructure which is present in the complex long events is unknown. Further investigation of the events within each category requires using features that can accurately capture the unique characteristics of each category and separate the events accordingly. The probability distributions for the event durations of the three different categories are shown in Figure 5.15A. The distributions of the short and long events overlap slightly between 50 and 200 μs, while the distributions of the long simple events and the long complex events overlap substantially from 50 μs to 1.2 ms. The overlap in distributions is quantified in Figure 5.15B, which shows cumulative probability distributions for event durations for a collection of short events and simple and complex long events which were manually selected from a single $I(t)$ recording. At 150 μs (indicated by the red vertical line on the left), 95% of short events are shorter in duration, while 60% of the simple long events and 15% of the complex long events are below this threshold. At 640 μs (indicated by the red vertical line on the right), 90% of simple events are shorter in duration, while 50% of the complex events are below this threshold, leading to some intermixing of simple and complex events in time.

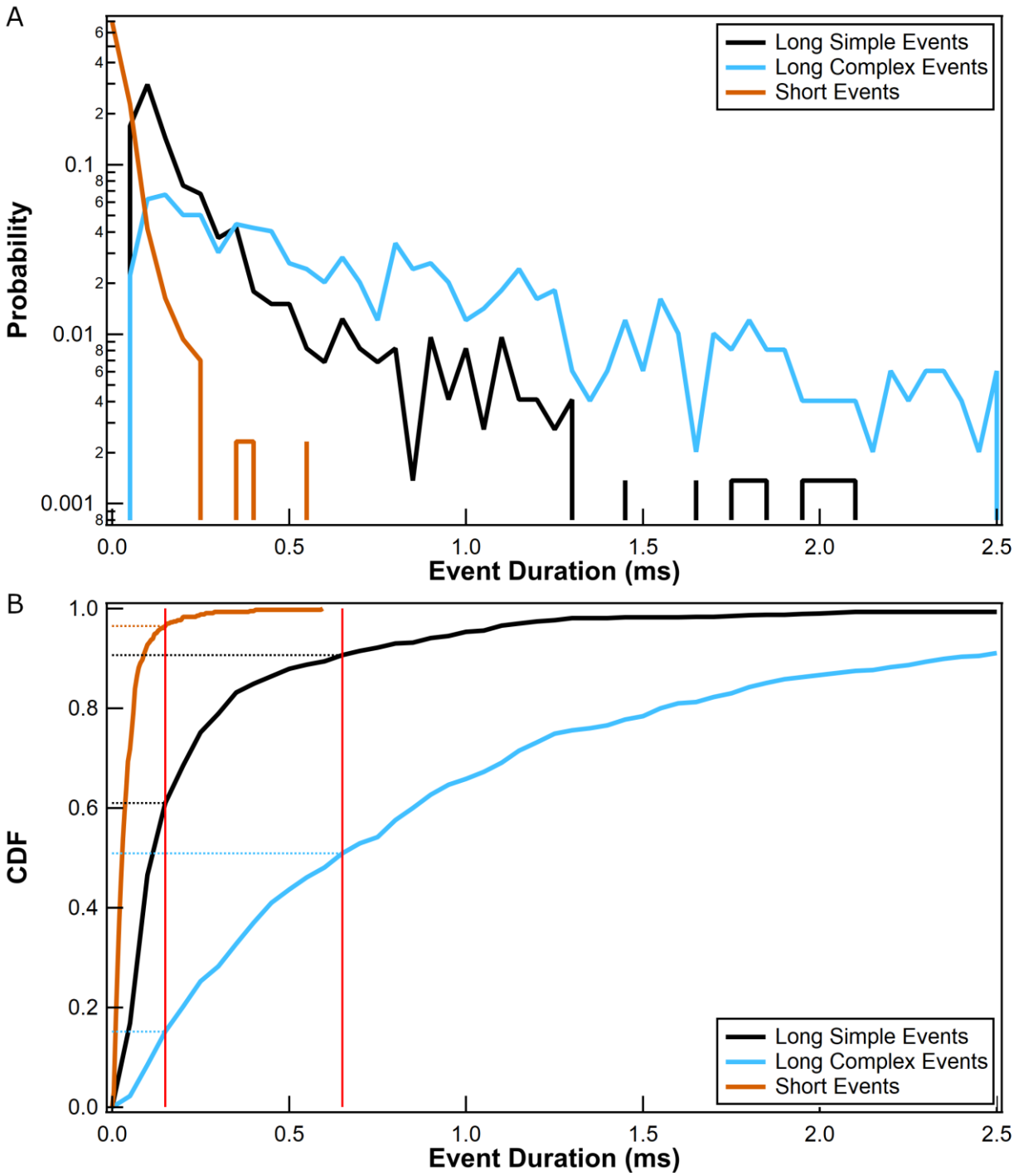

Figure 5.15: (A) Probability distributions and (B) cumulative probability distributions of the event durations for short events and simple and complex long events.

Although the short events can be mostly separated and identified by event duration, separating the long events into simple and complex categories requires a different approach.

Figure 5.16A shows the histograms of the denoised $I(t)$ total variation for both event types, which shows that the peaks of the two distributions can be separated at ~33 nA, but that there is still some overlap, especially between the tails of the distributions. The separation is more obvious when both features are compared in a scatter plot. Figure 5.16B shows a comparison of both features, where the simple events are plotted in black, and the complex events are plotted in light blue. The feature on the bottom axis is event duration, and the feature on the left is the total variation of the denoised $I(t)$ within each event. In this plot, events with denoised $I(t)$ total variation below ~4 nA are simple, while events above ~10 nA are complex. The dividing line between the simple and complex events is not a constant function with respect to either feature but, rather, a slowly-varying function of both denoised $I(t)$ total variation and event duration. The best-fit dividing line can be established with support vector machines (SVM), which is a supervised machine learning technique (199, 214).

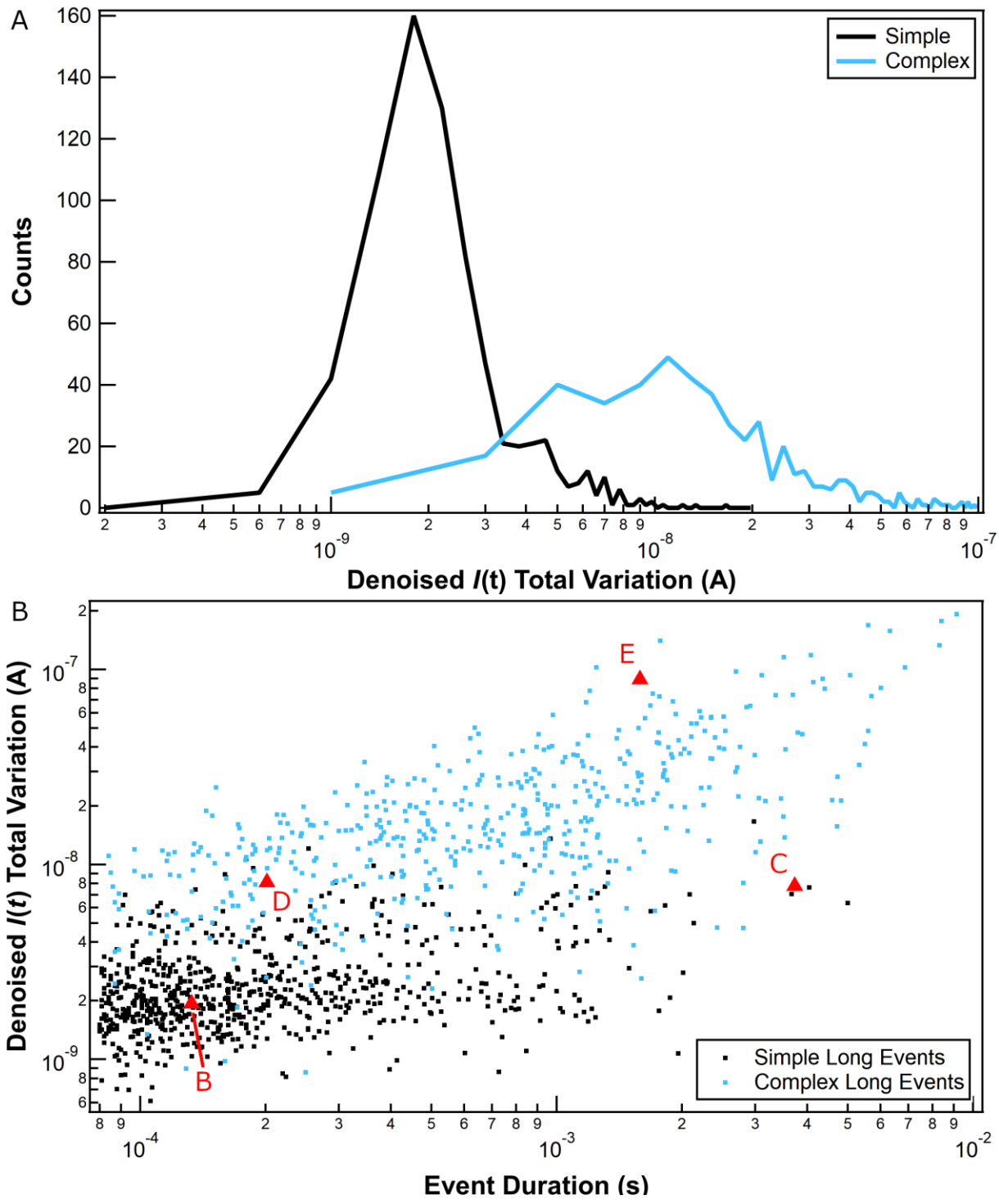

Figure 5.16: (A) Histograms of the denoised $I(t)$ total variation feature for simple and complex events, showing peaks that are shifted relative to each other. (B) Comparison two features of simple (black) and complex (light blue) long events, with the bottom axis corresponding to the duration of the event and the left axis corresponding to the total variation of the denoised $I(t)$ within an event. The points marked with red triangles correspond to the specific events illustrated in Figure 5.14.

Of the remaining features, the vast majority show no separation between simple and complex events, at least for characterizing catalytic activity of Taq DNA polymerase acting on $polyT_{42}$ and dATP.

5.6 Dimensionality Reduction and Principal Component Analysis of Event Features

Although previous sections demonstrate visual comparisons between two features, searching the entire feature set for significant correlations and separations requires another approach. Since each event is characterized by 189 individual features, it is essentially impossible to display more than two or three features simultaneously to perform visual comparisons. In addition, the most significant differences between two measurements might be best expressed as some combination of several features, which would be impossible to visualize if more than 3 features are involved. Fortunately, dimensionality reduction techniques can transform a dataset with many features into a reduced-dimension representation while still preserving the relationships in the original data (199). For example, many of the features in Figure 5.17, which shows the correlation matrix for events from both PCB48-KJ and PCB48-JK, show strong positive correlation (red and orange), indicating that those features exhibit similar trends and contain redundant information. The features exhibiting strong correlation, such as the event duration and the total variation in the UWT within an event, can be represented by some linear (or non-linear) combination of the individual features within the group, reducing several features down to one.

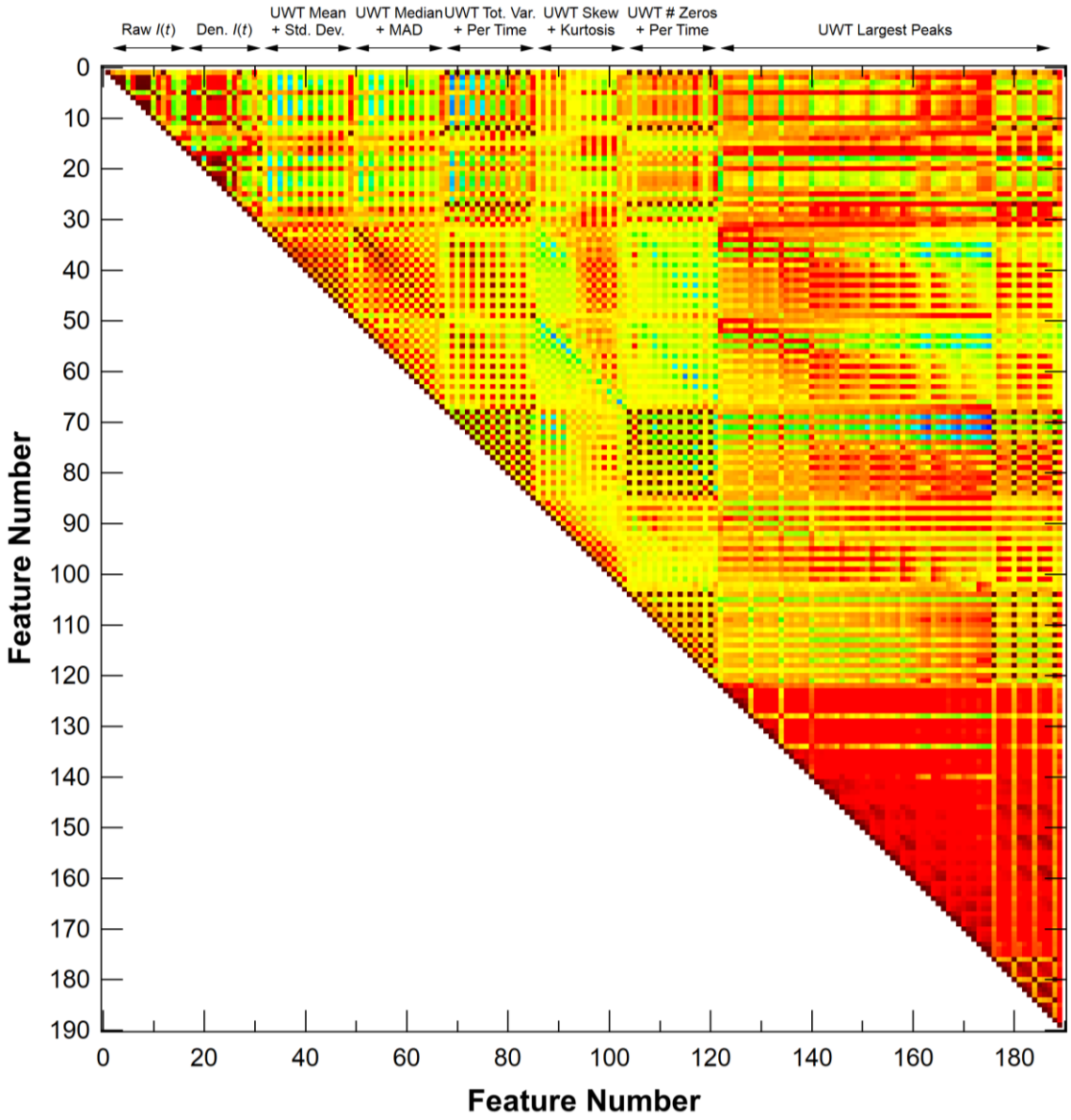

Figure 5.17: Correlation matrix showing correlations between all 189 features when the events from both PCB48-KJ and PCB48-JK are combined. Positive correlations are shown in dark red to orange, negative in green to blue, and no correlations in yellow.

Principal component analysis (PCA) is one basic method for dimensionality reduction (215). PCA looks for the principal components, or orthogonal directions in an N-dimensional dataset (where N is the number of features), which contain the most information, under the assumption that large variance indicates useful information (216). The method first

identifies the direction of maximal variance in all N dimensions and saves the direction as the first principal component. Then, it repeats the search for the direction of maximal variance in the (N-1)-dimensional subspace which is orthogonal to the first component, saves this second direction as the second principal component. The method continues searching for directions of maximal variance in the remaining subspaces (which are orthogonal to all the principal components already determined), until the variance in the principal components sums to 95% of the variance of the entire dataset.

Figure 5.18 shows the correlation matrix for the 65 principal components selected by PCA for the events in both PCB48-KJ and PCB48-JK. Since PCA is strongly influenced by the presence of multiple features which are strongly linearly correlated (multicollinear), any set of features with a correlation coefficient of more than 0.9 was reduced to a single feature before applying PCA. The pattern of correlations is much less structured than the pattern in the original features (Figure 5.17), but some correlations still exist between principal components. The strongest correlations (the darkest red and blue squares) involve the first 10 principal components, suggesting that those components contain the most important groupings of features. In fact, the features that contribute most to the first 10 principal components can be summarized by the labels listed in Table 5.2. As examples, lists of the most significant features within the first 5 principal components are given in Tables A.2 - A.6 in Appendix A. Some of these principal components contain features that are due to experimental artifacts rather than enzyme behavior (as discussed in Section 5.4), so removing those features may allow PCA to more effectively group biologically-important features together.

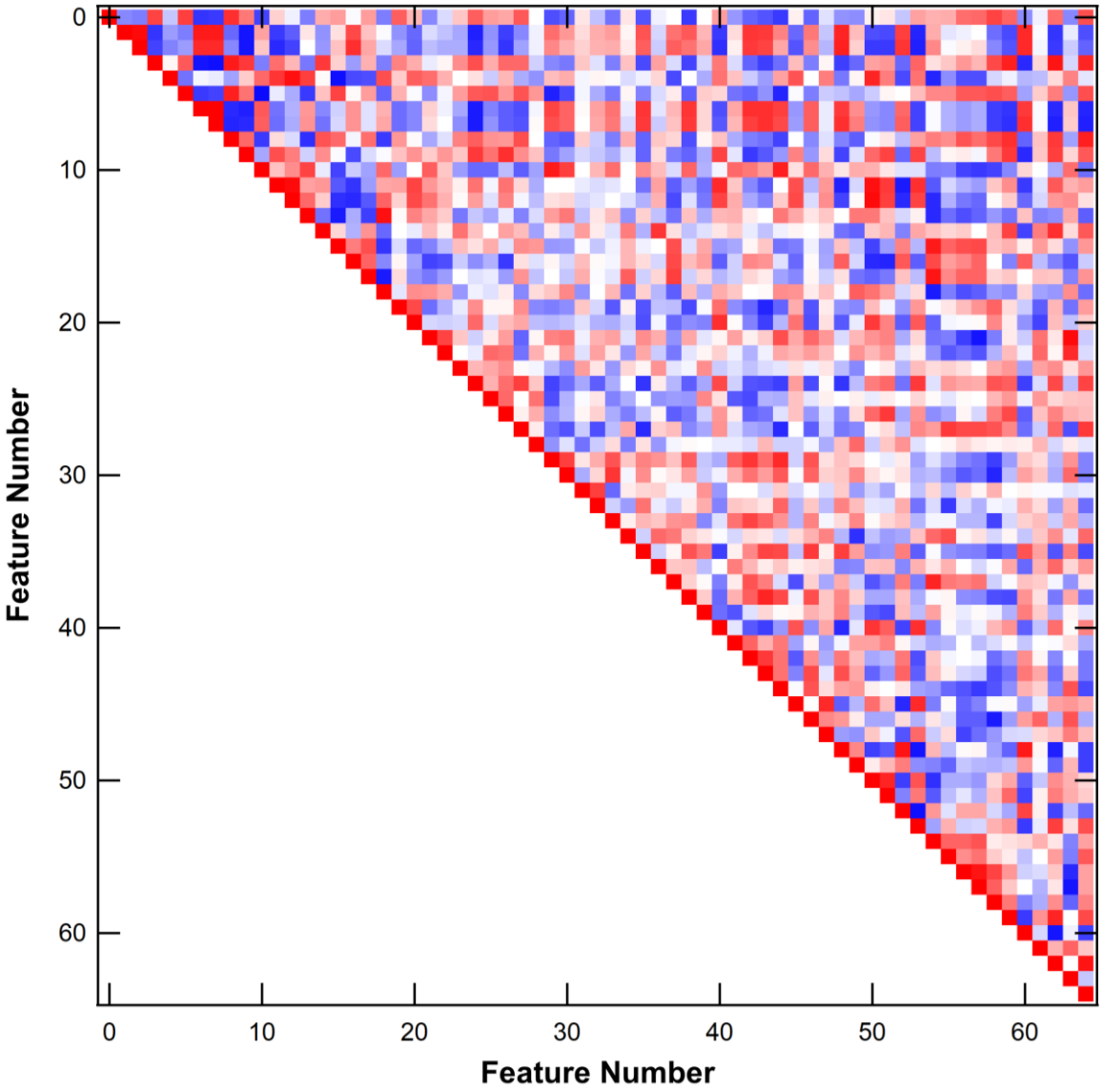

Figure 5.18: Correlation matrix showing correlations among 65 principal components, calculated by PCA, for events from both PCB48-KJ and PCB48-JK. Positive correlations are shown in red, negative correlations in blue, and no correlations in white.

Principal Component	Description of Dominant Features	Principal Component	Description of Dominant Features
1	UWT Peak Heights	6	I(t) Spread + UWT High-Frequency Spread
2	Event Duration	7	UWT Mid-Frequency Skewness + Kurtosis
3	Noise Strength	8	UWT Kurtosis + Spread
4	UWT Mid-Frequency Mean Values	9	UWT High-Frequency Noise + Distribution
5	Event + Noise Amplitudes	10	UWT High-Frequency Skewness + Mid-Frequency Spread

Table 5.2: List of the first 10 principal components, as identified by PCA for events from PCB48-KJ and PCB48-JK, and a summary of the main features within each component.

Since PCA uses linear transformations to find the principal components, it is unable to model any non-linear correlations between features. Though the principal components are linearly uncorrelated, some non-linear correlation may still exist between them. In addition, standard PCA works best on Gaussian-distributed data, but some of the event features, primarily those correlated with event duration, are exponentially distributed. Applying non-linear dimensionality-reduction techniques, such as kernel PCA (217) or sparse PCA (218), or techniques designed to handle exponential distributions (219-221), to the event features presented here may produce fewer principal components and a sparser representation. Further analysis can also utilize principal component regression, which explicitly looks for linear relationships between principal components rather than large variance (222).

5.7 Summary and Future Work

This chapter outlines an approach for analyzing biomolecule activity recorded in the $I(t)$ from a SWCNT-FET sensor by characterizing individual catalytic events rather than a collection of events. A fully-automated procedure for identifying individual events, including complex events, provides the infrastructure for extracting features from each event. When two features are compared, distributions of events are analyzed for correlations and separations of clusters. An example comparison between two measurements of Taq DNA polymerase show that the differences in feature distributions can result both from biologically-significant factors as well as biologically-insignificant noise or sensor variation. Finally, an initial analysis with PCA shows that the information contained in the 189 initial features can be captured with 65 vectors, which is 1/3 of the size of the original feature set.

This chapter only discusses the features corresponding to a single type of complementary nucleotide incorporation (poly T_{42} and dATP) by Taq DNA polymerase. The incorporation of other nucleotide bases to their complementary templates, or even the incorporation of multiple nucleotide bases on the same template, may exhibit correlation or clustering patterns that differ from those shown here for poly T_{42} and dATP. When the analyses described in this chapter are applied to various combinations of nucleotides and DNA templates, the signal differences between the various combinations can be quantified using the event features.

Future work will include applying machine learning techniques to the state vectors corresponding to individual events, which can assist with several tasks. The first task is identifying the significant features and characteristic values that correspond to specific nucleotide bases, DNA templates or mutants. This can be done with self-supervised learning techniques such as autoencoders (223, 224). The second task is identifying which features and values of those features contribute to accurate base calling during nucleotide incorporation. Supervised learning techniques like support vector machines (SVMs) can determine the most significant collection of features by analyzing a pre-labeled training set of various nucleotide and template combinations (199, 214, 225). SVMs can also find the hyperplane that best separates events by nucleotide identity. This second task is essential for developing accurate base calling, which is a prerequisite for applying the SWCNT-FET biosensor to DNA sequencing.

REFERENCES

1. Choi YK, Moody IS, Sims PC, Hunt SR, Corso BL, Perez I, et al. Single-Molecule Lysozyme Dynamics Monitored by an Electronic Circuit. Science. 2012;335(6066):319-24.

2. Olsen TJ, Choi Y, Sims PC, Gu OT, Corso BL, Dong CJ, et al. Electronic Measurements of Single-Molecule Processing by DNA Polymerase I (Klenow Fragment). Journal of the American Chemical Society. 2013;135(21):7855-60.

3. Sims PC, Moody IS, Choi Y, Dong CJ, Iftikhar M, Corso BL, et al. Electronic Measurements of Single-Molecule Catalysis by cAMP-Dependent Protein Kinase A. Journal of the American Chemical Society. 2013;135(21):7861-8.

4. Peterman EJ, Wuite GJ. Single Molecule Analysis: Springer; 2018.

5. Miller H, Zhou ZK, Shepherd J, Wollman AJM, Leake MC. Single-molecule techniques in biophysics: a review of the progress in methods and applications. Rep Prog Phys. 2018;81(2).

6. Gao DL, Ding WQ, Nieto-Vesperinas M, Ding XM, Rahman M, Zhang TH, et al. Optical manipulation from the microscale to the nanoscale: fundamentals, advances and prospects. Light-Sci Appl. 2017;6.

7. Fazal FM, Block SM. Optical tweezers study life under tension. Nat Photonics. 2011;5(6):318-21.

8. Gross P, Farge G, Peterman EJG, Wuite GJL. Combining Optical Tweezers, Single-Molecule Fluorescence Microscopy, and Microfluidics for Studies of DNA-Protein Interactions. Methods in Enzymology, Vol 475: Single Molecule Tools, Pt B. 2010;475:427-53.

9. Capitanio M, Pavone FS. Interrogating Biology with Force: Single Molecule High-Resolution Measurements with Optical Tweezers. Biophys J. 2013;105(6):1293-303.

10. Heller I, Hoekstra TP, King GA, Peterman EJG, Wuite GJL. Optical Tweezers Analysis of DNA-Protein Complexes. Chem Rev. 2014;114(6):3087-119.

11. Choudhary D, Mossa A, Jadhav M, Cecconi C. Bio-Molecular Applications of Recent Developments in Optical Tweezers. Biomolecules. 2019;9(1).

12. Holmstrom ED, Holla A, Zheng WW, Nettels D, Best RB, Schuler B. Accurate Transfer Efficiencies, Distance Distributions, and Ensembles of Unfolded and Intrinsically Disordered Proteins From Single-Molecule FRET. Intrinsically Disordered Proteins. 2018;611:287-325.

13. Lerner E, Cordes T, Ingargiola A, Alhadid Y, Chung S, Michalet X, et al. Toward dynamic structural biology: Two decades of single-molecule Forster resonance energy transfer. Science. 2018;359(6373):288-+.

14. Blanco M, Walter NG. Analysis of Complex Single-Molecule Fret Time Trajectories. Methods in Enzymology, Vol 472: Single Molecule Tools, Pt A: Fluorescence Based Approaches. 2010;472:153-78.

15. Salem CB, Ploetz E, Lamb DC. Probing dynamics in single molecules. Devl Phys Th Chem. 2019:71-115.

16. Hwang H, Kim H, Myong S. Protein induced fluorescence enhancement as a single molecule assay with short distance sensitivity. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(18):7414-8.

17. Stennett EMS, Ciuba MA, Lin S, Levitus M. Demystifying PIFE: The Photophysics Behind the Protein-Induced Fluorescence Enhancement Phenomenon in Cy3. J Phys Chem Lett. 2015;6(10):1819-23.

18. Zhang JD, Chi QJ, Hansen AG, Jensen PS, Salvatore P, Ulstrup J. Interfacial electrochemical electron transfer in biology - Towards the level of the single molecule. Febs Lett. 2012;586(5):526-35.

19. Zhang B, Song W, Brown J, Nemanich R, Lindsay S. Electronic Conductance Resonance in Non-Redox-Active Proteins. J Am Chem Soc. 2020;142(13):6432-8.

20. Zhao YA, Ashcroft B, Zhang PM, Liu H, Sen SM, Song W, et al. Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. Nat Nanotechnol. 2014;9(6):466-73.

21. Im J, Biswas S, Liu H, Zhao YN, Sen SM, Biswas S, et al. Electronic single-molecule identification of carbohydrate isomers by recognition tunnelling. Nat Commun. 2016;7.

22. Im J, Sen S, Lindsay S, Zhang PM. Recognition Tunneling of Canonical and Modified RNA Nucleotides for Their Identification with the Aid of Machine Learning. Acs Nano. 2018;12(7):7067-75.

23. Willems K, Van Meervelt V, Wloka C, Maglia G. Single-molecule nanopore enzymology. Philos T R Soc B. 2017;372(1726).

24. Varongchayakul N, Song JX, Meller A, Grinstaff MW. Single-molecule protein sensing in a nanopore: a tutorial. Chemical Society Reviews. 2018;47(23).

25. Choi Y, Olsen TJ, Sims PC, Moody IS, Corso BL, Dang MN, et al. Dissecting Single-Molecule Signal Transduction in Carbon Nanotube Circuits with Protein Engineering. Nano Letters. 2013;13(2):625-31.

26. Akhterov MV, Choi Y, Olsen TJ, Sims PC, Iftikhar M, Gul OT, et al. Observing Lysozyme's Closing and Opening Motions by High-Resolution Single-Molecule Enzymology. Acs Chem Biol. 2015;10(6):1495-501.

27. Pacios M, Martin-Fernandez I, Borrise X, del Valle M, Bartroli J, Lora-Tamayo E, et al. Real time protein recognition in a liquid-gated carbon nanotube field-effect transistor modified with aptamers. Nanoscale. 2012;4(19):5917-23.

28. Lerner MB, Matsunaga F, Han GH, Hong SJ, Xi J, Crook A, et al. Scalable Production of Highly Sensitive Nanosensors Based on Graphene Functionalized with a Designed G Protein-Coupled Receptor. Nano Letters. 2014;14(5):2709-14.

29. Collins PG, Fuhrer MS, Zettl A. 1/f noise in carbon nanotubes. Applied Physics Letters. 2000;76(7):894-6.

30. Liu F, Bao MQ, Kim HJ, Wang KL, Li C, Liu XL, et al. Giant random telegraph signals in the carbon nanotubes as a single defect probe. Applied Physics Letters. 2005;86(16).

31. Liu F, Wang KL, Zhang DH, Zhou CW. Random telegraph signals and noise behaviors in carbon nanotube transistors. Applied Physics Letters. 2006;89(24).

32. Vishnubhotla R, Ping JL, Gao ZL, Lee A, Saouaf O, Vrudhula A, et al. Scalable graphene aptasensors for drug quantification. Aip Advances. 2017;7(11).

33. Kim YG, Ho SO, Gassman NR, Korlann Y, Landorf EV, Collart FR, et al. Efficient site-specific Labeling of proteins via cysteines. Bioconjugate Chemistry. 2008;19(3):786-91.

34. Koniev O, Wagner A. Developments and recent advancements in the field of endogenous amino acid selective bond forming reactions for bioconjugation. Chemical Society Reviews. 2015;44(15):5495-551.

35. Johnson KA. The kinetic and chemical mechanism of high-fidelity DNA polymerases. Bba-Proteins Proteom. 2010;1804(5):1041-8.

36. Joyce CM. Techniques used to study the DNA polymerase reaction pathway. Bba-Proteins Proteom. 2010;1804(5):1032-40.

37. Yang SG, Cao JS, Silbey RJ, Sung JY. Quantitative Interpretation of the Randomness in Single Enzyme Turnover Times. Biophys J. 2011;101(3):519-24.

38. Liebovitch LS, Toth TI. DISTRIBUTIONS OF ACTIVATION-ENERGY BARRIERS THAT PRODUCE STRETCHED EXPONENTIAL PROBABILITY-DISTRIBUTIONS FOR THE TIME SPENT IN EACH STATE OF THE 2 STATE REACTION A-REVERSIBLE-B. Bulletin of Mathematical Biology. 1991;53(3):443-55.

39. English BP, Min W, van Oijen AM, Lee KT, Luo GB, Sun HY, et al. Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. Nat Chem Biol. 2006;2(2):87-94.

40. Dan N. Understanding dynamic disorder fluctuations in single-molecule enzymatic reactions. Curr Opin Colloid In. 2007;12(6):314-21.

41. Holford TR, Davis F, Higson SP. Recent trends in antibody based sensors. Biosensors and Bioelectronics. 2012;34(1):12-24.

42. Yang XD, Corvalan JR, Wang P, Roy CMN, Davis CG. Fully human anti-interleukin-8 monoclonal antibodies: potential therapeutics for the treatment of inflammatory disease states. Journal of Leukocyte Biology. 1999;66(3):401-10.

43. Engvall E. [28] Enzyme immunoassay ELISA and EMIT. Methods in enzymology. 70: Elsevier; 1980. p. 419-39.

44. Conroy PJ, Hearty S, Leonard P, O'Kennedy RJ, editors. Antibody production, design and use for biosensor-based applications. Seminars in cell & developmental biology; 2009: Elsevier.

45. Skottrup PD, Nicolaisen M, Justesen AF. Towards on-site pathogen detection using antibody-based sensors. Biosensors and Bioelectronics. 2008;24(3):339-48.

46. Afsahi S, Lerner MB, Goldstein JM, Lee J, Tang X, Bagarozzi Jr DA, et al. Novel graphene-based biosensor for early detection of Zika virus infection. Biosensors and Bioelectronics. 2018;100:85-8.

47. Klein J, Nikolaidis N. The descent of the antibody-based immune system by gradual evolution. Proceedings of the National Academy of Sciences. 2005;102(1):169-74.

48. Wang W, Singh S, Zeng DL, King K, Nema S. Antibody structure, instability, and formulation. Journal of pharmaceutical sciences. 2007;96(1):1-26.

49. Braden BC, Dall'Acqua W, Eisenstein E, Fields BA, Goldbaum FA, Malchiodi EL, et al. Protein motion and lock and key complementarity in antigen-antibody reactions. Pharm Acta Helv. 1995;69(4):225-30.

50. James LC, Roversi P, Tawfik DS. Antibody multispecificity mediated by conformational diversity. Science. 2003;299(5611):1362-7.

51. Notkins AL. Polyreactivity of antibody molecules. Trends Immunol. 2004;25(4):174-9.

52. Organization WH. WHO Model List of Essential Medicines, 20th list, March 2017. World Health Organization; 2017.

53. Gelderblom H, Mross K, ten Tije AJ, Behringer D, Mielke S, van Zomeren DM, et al. Comparative pharmacokinetics of unbound paclitaxel during 1-and 3-hour infusions. Journal of Clinical Oncology. 2002;20(2):574-81.

54. Sparreboom A, van Tellingen O, Nooijen WJ, Beijnen JH. Nonlinear pharmacokinetics of paclitaxel in mice results from the pharmaceutical vehicle Cremophor EL. Cancer research. 1996;56(9):2112-5.

55. Wenk MR, Fahr A, Reszka R, Seelig J. Paclitaxel partitioning into lipid bilayers. Journal of pharmaceutical sciences. 1996;85(2):228-31.

56. van Zuylen L, Verweij J, Sparreboom A. Role of formulation vehicles in taxane pharmacology. Investigational New Drugs. 2001;19(2):125-41.

57. Scheithauer W, Ramanathan RK, Moore M, Macarulla T, Goldstein D, Hammel P, et al. Dose modification and efficacy of nab-paclitaxel plus gemcitabine vs. gemcitabine for patients with metastatic pancreatic cancer: phase III MPACT trial. Journal of gastrointestinal oncology. 2016;7(3):469.

58. Kampan NC, Madondo MT, McNally OM, Quinn M, Plebanski M. Paclitaxel and its evolving role in the management of ovarian cancer. BioMed research international. 2015;2015.

59. Bignami GS, Mooberry SL. Monoclonal antibodies to taxanes that neutralize the biological activity of paclitaxel. Cancer Letters. 1998;126(2):127-33.

60. Lee YJ, Park C. Methods, devices, and reagents for monitoring paclitaxel concentration in plasma for pharmacokinetic-guided dosing of paclitaxel. Google Patents; 2018.

61. Marx V. Finding the right antibody for the job. Nature Methods. 2013;10(8):703-7.

62. Voskuil JLA. The challenges with the validation of research antibodies. F1000Res. 2017;6:161-.

63. Beuwer MA, Prins MWJ, Zijlstra P. Stochastic Protein Interactions Monitored by Hundreds of Single-Molecule Plasmonic Biosensors. Nano Letters. 2015;15(5):3507-11.

64. Coppari E, Santini S, Bizzarri AR, Cannistraro S. Kinetics and binding geometries of the complex between beta(2)-microglobulin and its antibody: An AFM and SPR study. Biophysical Chemistry. 2016;211:19-27.

65. Patel R, Andrien BA. Kinetic analysis of a monoclonal therapeutic antibody and its single-chain homolog by surface plasmon resonance. Analytical Biochemistry. 2010;396(1):59-68.

66. Heinrich L, Tissot N, Hartmann DJ, Cohen R. Comparison of the results obtained by ELISA and surface plasmon resonance for the determination of antibody affinity. Journal of Immunological Methods. 2010;352(1-2):13-22.

67. Bantz KC, Meyer AF, Wittenberg NJ, Im H, Kurtulus O, Lee SH, et al. Recent progress in SERS biosensing. Physical Chemistry Chemical Physics. 2011;13(24):11551-67.

68. Sasakawa H, Sakata E, Yamaguchi Y, Masuda M, Mori T, Kurimoto E, et al. Ultra-high field NMR studies of antibody binding and site-specific phosphorylation of alpha-synuclein. Biochemical and Biophysical Research Communications. 2007;363(3):795-9.

69. Lee CK, Wang YM, Huang LS, Lin SM. Atomic force microscopy: Determination of unbinding force, off rate and energy barrier for protein-ligand interaction. Micron. 2007;38(5):446-61.

70. Kelliher MT, Jacks RD, Piraino MS, Southern CA. The effect of sugar removal on the structure of the Fc region of an IgG antibody as observed with single molecule Forster Resonance Energy Transfer. Molecular Immunology. 2014;60(2):103-8.

71. Bhunia D, Chowdhury R, Bhattacharyya K, Ghosh S. Fluorescence fluctuation of an antigen-antibody complex: circular dichroism, FCS and smFRET of enhanced GFP and its antibody. Physical Chemistry Chemical Physics. 2015;17(38):25250-9.

72. Tian J, Stella VJ. Degradation of paclitaxel and related compounds in aqueous solutions II: Nonepimerization degradation under neutral to basic pH conditions. Journal of Pharmaceutical Sciences. 2008;97(8):3100-8.

73. Lerner MB, D'Souza J, Pazina T, Dailey J, Goldsmith BR, Robinson MK, et al. Hybrids of a Genetically Engineered Antibody and a Carbon Nanotube Transistor for Detection of Prostate Cancer Biomarkers. Acs Nano. 2012;6(6):5143-9.

74. Topinka MA, Rowell MW, Goldhaber-Gordon D, McGehee MD, Hecht DS, Gruner G. Charge Transport in Interpenetrating Networks of Semiconducting and Metallic Carbon Nanotubes. Nano Letters. 2009;9(5):1866-71.

75. Reuel NF, Bojo P, Zhang JQ, Boghossian AA, Ahn JH, Kim JH, et al. NoRSE: noise reduction and state evaluator for high-frequency single event traces. Bioinformatics. 2012;28(2):296-7.

76. Gevondyan NM, Volynskaia AM, Gevondyan VS. Four free cysteine residues found in human IgG1 of healthy donors. Biochemistry-Moscow. 2006;71(3):279-84.

77. Liu H, May K, editors. Disulfide bond structures of IgG molecules: structural variations, chemical modifications and possible impacts to stability and biological function. MAbs; 2012: Taylor & Francis.

78. Harris LJ, Skaletsky E, McPherson A. Crystallographic structure of an intact IgG1 monoclonal antibody11Edited by I. A. Wilson. Journal of Molecular Biology. 1998;275(5):861-72.

79. Chumsae C, Gaza-Bulseco G, Liu HC. Identification and Localization of Unpaired Cysteine Residues in Monoclonal Antibodies by Fluorescence Labeling and Mass Spectrometry. Anal Chem. 2009;81(15):6449-57.

80. Kulin S, Kishore R, Hubbard JB, Helmerson K. Real-time measurement of spontaneous antigen-antibody dissociation. Biophys J. 2002;83(4):1965-73.

81. Maxwell BA, Suo ZC. Single-molecule Investigation of Substrate Binding Kinetics and Protein Conformational Dynamics of a B-family Replicative DNA Polymerase. Journal of Biological Chemistry. 2013;288(16):11590-600.

82. Goldsmith BR, Mitala JJ, Josue J, Castro A, Lerner MB, Bayburt TH, et al. Biomimetic Chemical Sensors Using Nanoelectronic Readout of Olfactory Receptor Proteins. Acs Nano. 2011;5(7):5408-16.

83. Schwesinger F, Ros R, Strunz T, Anselmetti D, Guntherodt HJ, Honegger A, et al. Unbinding forces of single antibody-antigen complexes correlate with their thermal dissociation rates. Proceedings of the National Academy of Sciences of the United States of America. 2000;97(18):9972-7.

84. Yin LL, Yang YZ, Wang SP, Wang W, Zhang ST, Tao NJ. Measuring Binding Kinetics of Antibody-Conjugated Gold Nanoparticles with Intact Cells. Small. 2015;11(31):3782-8.

85. Blake RC, Delehanty JB, Khosraviani M, Yu H, Jones RM, Blake DA. Allosteric binding properties of a monoclonal antibody and its Fab fragment. Biochemistry. 2003;42(2):497-508.

86. Cannon B, Weaver N, Pu Q, Thiagarajan V, Liu S, Huang J, et al. Cholesterol modulated antibody binding in supported lipid membranes as determined by total internal reflectance microscopy on a microfabricated high-throughput glass chip. Langmuir. 2005;21(21):9666-74.

87. Hattori T, Lai D, Dementieva IS, Montano SP, Kurosawa K, Zheng YP, et al. Antigen clasping by two antigen-binding sites of an exceptionally specific antibody for histone methylation. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(8):2092-7.

88. Genna V, Donati E, De Vivo M. The Catalytic Mechanism of DNA and RNA Polymerases. Acs Catal. 2018;8(12):11103-18.

89. Wu WJ, Yang W, Tsai MD. How DNA polymerases catalyse replication and repair with contrasting fidelity. Nat Rev Chem. 2017;1(9).

90. Reha-Krantz LJ. DNA polymerase proofreading: Multiple roles maintain genome stability. Bba-Proteins Proteom. 2010;1804(5):1049-63.

91. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 2009;323(5910):133-8.

92. Chen CY. DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. Front Microbiol. 2014;5.

93. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333-51.

94. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017;550(7676):345-53.

95. Mendez J, Blanco L, Lazaro JM, Salas M. Primer-Terminus Stabilization at the Phi-29 DNA-Polymerase Active-Site - Mutational Analysis of Conserved Motif Tx(2)Gr. Journal of Biological Chemistry. 1994;269(47):30030-8.

96. deVega M, Lazaro JM, Salas M, Blanco L. Primer terminus stabilization at the 3'-5' exonuclease active site of phi 29 DNA polymerase. Involvement of two amino acid residues highly conserved in proofreading DNA polymerases. Embo J. 1996;15(5):1182-92.

97. Bonnin A, Lazaro JM, Blanco L, Salas M. A single tyrosine prevents insertion of ribonucleotides in the eukaryotic-type phi 29 DNA polymerase. Journal of Molecular Biology. 1999;290(1):241-51.

98. Suzuki M, Yoshida S, Adman ET, Blank A, Loeb LA. Thermus aquaticus DNA polymerase I mutants with altered fidelity - Interacting mutations in the O-helix. Journal of Biological Chemistry. 2000;275(42):32728-35.

99. Yoshida K, Tosaka A, Kamiya H, Murate T, Kasai H, Nimura Y, et al. Arg660Ser mutation in Thermus aquaticus DNA polymerase I suppresses T -> C transitions: implication of wobble base pair formation at the nucleotide incorporation step. Nucleic Acids Res. 2001;29(20):4206-14.

100. Rodriguez I, Lazaro JM, Salas M, de Vega M. phi 29 DNA polymerase-terminal protein interaction. Involvement of residues specifically conserved among protein-primed DNA polymerases. Journal of Molecular Biology. 2004;337(4):829-41.

101. Rodriguez I, Lazaro JM, Blanco L, Kamtekar S, Berman AJ, Wang JM, et al. A specific subdomain in phi 29 DNA polymerase confers both processivity and strand-displacement capacity. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(18):6407-12.

102. Li Y, Korolev S, Waksman G. Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of Thermus aquaticus DNA polymerase I: structural basis for nucleotide incorporation. Embo J. 1998;17(24):7514-25.

103. Kamtekar S, Berman AJ, Wang JM, Lazaro JM, de Vega M, Blanco L, et al. Insights into strand displacement and processivity from the crystal structure of the protein-primed DNA polymerase of bacteriophage phi 29. Mol Cell. 2004;16(4):609-18.

104. Berman AJ, Kamtekar S, Goodman JL, Lazaro JM, de Vega M, Blanco L, et al. Structures of phi29 DNA polymerase complexed with substrate: the mechanism of translocation in B-family polymerases. Embo J. 2007;26(14):3494-505.

105. Rothwell PJ, Mitaksov V, Waksman G. Motions of the fingers subdomain of klentaq1 are fast and not rate limiting: Implications for the molecular basis of fidelity in DNA polymerases. Mol Cell. 2005;19(3):345-55.

106. Xu CL, Maxwell BA, Suo ZC. Conformational Dynamics of Thermus aquaticus DNA Polymerase I during Catalysis. Journal of Molecular Biology. 2014;426(16):2901-17.

107. Garmendia C, Bernad A, Esteban JA, Blanco L, Salas M. The Bacteriophage-Phi-29 DNA-Polymerase, a Proofreading Enzyme. Journal of Biological Chemistry. 1992;267(4):2594-9.

108. Esteban JA, Salas M, Blanco L. Fidelity of Phi-29 DNA-Polymerase - Comparison between Protein-Primed Initiation and DNA Polymerization. Journal of Biological Chemistry. 1993;268(4):2719-26.

109. Soengas MS, Gutierrez C, Salas M. Helix-Destabilizing Activity of Phi-29 Single-Stranded-DNA Binding-Protein - Effect on the Elongation Rate during Strand Displacement DNA-Replication. Journal of Molecular Biology. 1995;253(4):517-29.

110. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification. Genome Res. 2001;11(6):1095-9.

111. Ibarra B, Chemla YR, Plyasunov S, Smith SB, Lazaro JM, Salas M, et al. Proofreading dynamics of a processive DNA polymerase. Embo J. 2009;28(18):2794-802.

112. Morin JA, Cao FJ, Lazaro JM, Arias-Gonzalez JR, Valpuesta JM, Carrascosa JL, et al. Active DNA unwinding dynamics during processive DNA replication. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(21):8115-20.

113. Morin JA, Cao FJ, Valpuesta JM, Carrascosa JL, Salas M, Ibarra B. Manipulation of single polymerase-DNA complexes A mechanical view of DNA unwinding during replication. Cell Cycle. 2012;11(16):2967-8.

114. Morin JA, Cao FJ, Lazaro JM, Arias-Gonzalez JR, Valpuesta JM, Carrascosa JL, et al. Mechano-chemical kinetics of DNA replication: identification of the translocation step of a replicative DNA polymerase. Nucleic Acids Res. 2015;43(7):3643-52.

115. Lieberman KR, Cherf GM, Doody MJ, Olasagasti F, Kolodji Y, Akeson M. Processive Replication of Single DNA Molecules in a Nanopore Catalyzed by phi29 DNA Polymerase. Journal of the American Chemical Society. 2010;132(50):17961-72.

116. Manrao EA, Derrington IM, Laszlo AH, Langford KW, Hopper MK, Gillgren N, et al. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. Nat Biotechnol. 2012;30(4):349-U174.

117. Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M. Automated forward and reverse ratcheting of DNA in a nanopore at 5-angstrom precision. Nat Biotechnol. 2012;30(4):344-8.

118. Dahl JM, Mai AH, Cherf GM, Jetha NN, Garalde DR, Marziali A, et al. Direct Observation of Translocation in Individual DNA Polymerase Complexes. Journal of Biological Chemistry. 2012;287(16):13407-21.

119. Lieberman KR, Dahl JM, Wang HY. Kinetic Mechanism at the Branchpoint between the DNA Synthesis and Editing Pathways in Individual DNA Polymerase Complexes. Journal of the American Chemical Society. 2014;136(19):7117-31.

120. Dahl JM, Wang HY, Lazaro JM, Salas M, Lieberman KR. Kinetic Mechanisms Governing Stable Ribonucleotide Incorporation in Individual DNA Polymerase Complexes. Biochemistry. 2014;53(51):8061-76.

121. Schwartz JJ, Quake SR. Single molecule measurement of the "speed limit'' of DNA polymerase. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(48):20294-9.

122. de Vega M, Blanco L, Salas M. Processive proofreading and the spatial relationship between polymerase and exonuclease active sites of bacteriophage o29 DNA polymerase. Journal of Molecular Biology. 1999;292(1):39-51.

123. Blanco L, Bernad A, Lazaro JM, Martin G, Garmendia C, Salas M. Highly Efficient DNA-Synthesis by the Phage Phi-29 DNA-Polymerase - Symmetrical Mode of DNA-Replication. Journal of Biological Chemistry. 1989;264(15):8935-40.

124. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009;4(4):265-70.

125. Mantere T, Kersten S, Hoischen A. Long-Read Sequencing Emerging in Medical Genetics. Front Genet. 2019;10.

126. Lieberman KR, Dahl JM, Mai AH, Cox A, Akeson M, Wang HY. Kinetic Mechanism of Translocation and dNTP Binding in Individual DNA Polymerase Complexes. Journal of the American Chemical Society. 2013;135(24):9149-55.

127. Pugliese KM, Gul OT, Choi Y, Olsen TJ, Sims PC, Collins PG, et al. Processive Incorporation of Deoxynucleoside Triphosphate Analogs by Single-Molecule DNA Polymerase I (Klenow Fragment) Nanocircuits. Journal of the American Chemical Society. 2015;137(30):9587-94.

128. Joyce CM, Potapova O, DeLucia AM, Huang XW, Basu VP, Grindley NDF. Fingers-closing and other rapid conformational changes in DNA polymerase I (Klenow fragment) and their role in nucleotide selectivity. Biochemistry. 2008;47(23):6103-16.

129. Henke W, Herdel K, Jung K, Schnorr D, Loening SA. Betaine improves the PCR amplification of GC-rich DNA sequences. Nucleic Acids Res. 1997;25(19):3957-8.

130. Zhang ZZ, Yang X, Meng LY, Liu F, Shen CC, Yang WJ. Enhanced amplification of GC-rich DNA with two organic reagents. Biotechniques. 2009;47(3):775-8.

131. Largy E, Mergny J-L, Gabelica V. Role of alkali metal ions in G-quadruplex nucleic acid structure and stability. The Alkali Metal Ions: Their Role for Life: Springer; 2016. p. 203-58.

132. Spiegel J, Adhikari S, Balasubramanian S. The Structure and Function of DNA G-Quadruplexes. Trends Chem. 2020;2(2):123-36.

133. Hagerman PJ. Flexibility of DNA. Annu Rev Biophys Bio. 1988;17:265-86.

134. Mills JB, Vacano E, Hagerman PJ. Flexibility of single-stranded DNA: Use of gapped duplex helices to determine the persistence lengths of poly(dT) and poly(dA). Journal of Molecular Biology. 1999;285(1):245-57.

135. Murphy MC, Rasnik I, Cheng W, Lohman TM, Ha TJ. Probing single-stranded DNA conformational flexibility using fluorescence spectroscopy. Biophys J. 2004;86(4):2530-7.

136. Chen H, Meisburger SP, Pabit SA, Sutton JL, Webb WW, Pollack L. Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(3):799-804.

137. Yu G, Mallat S, Bacry E. Audio denoising by time-frequency block thresholding. Ieee T Signal Proces. 2008;56(5):1830-9.

138. Chambolle A, DeVore RA, Lee NY, Lucier BJ. Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage. Ieee T Image Process. 1998;7(3):319-35.

139. Portilla J, Strela V, Wainwright MJ, Simoncelli EP. Image denoising using scale mixtures of Gaussians in the wavelet domain. Ieee T Image Process. 2003;12(11):1338-51.

140. Hou R, Wang ZY, Diamond JJ, Zheng Z, Zhu JS, Wang ZC, et al. A quantitative evaluation model of denoising methods for surface plasmon resonance imaging signal. Sensor Actuat B-Chem. 2011;160(1):951-6.

141. AlMahamdy M, Riley HB. Performance Study of Different Denoising Methods for ECG Signals. Procedia Comput Sci. 2014;37:325-+.

142. Vazquez RR, Velez-Perez H, Ranta R, Dorr VL, Maquin D, Maillard L. Blind source separation, wavelet denoising and discriminant analysis for EEG artefacts and noise cancelling. Biomed Signal Proces. 2012;7(4):389-400.

143. Liu YP, Li Y, Lin HB, Ma HT. An Amplitude-Preserved Time-Frequency Peak Filtering Based on Empirical Mode Decomposition for Seismic Random Noise Reduction. Ieee Geosci Remote S. 2014;11(5):896-900.

144. To AC, Moore JR, Glaser SD. Wavelet denoising techniques with applications to experimental geophysical data. Signal Process. 2009;89(2):144-60.

145. Wang JH, Lin LD. Improved median filter using minmax algorithm for image processing. Electron Lett. 1997;33(16):1362-3.

146. Chambolle A. An algorithm for total variation minimization and applications. J Math Imaging Vis. 2004;20(1-2):89-97.

147. Candan C, Inan H. A unified framework for derivation and implementation of Savitzky-Golay filters. Signal Process. 2014;104:203-11.

148. Savitzky A, Golay MJE. Smoothing + Differentiation of Data by Simplified Least Squares Procedures. Anal Chem. 1964;36(8):1627-&.

149. Mallat S, Peyre G. A Wavelet Tour of Signal Processing The Sparse Way Preface to the Sparse Edition. Wavelet Tour of Signal Processing: The Sparse Way. 2009:Xv-+.

150. Kauppinen I, Roth K. Improved noise reduction in audio signals using spectral resolution enhancement with time-domain signal extrapolation. Ieee T Speech Audi P. 2005;13(6):1210-6.

151. Buades A, Coll B, Morel JM. A review of image denoising algorithms, with a new one. Multiscale Model Sim. 2005;4(2):490-530.

152. Balster EJ, Zheng YF, Ewing RL. Feature-based wavelet shrinkage algorithm for image denoising. Ieee T Image Process. 2005;14(12):2024-39.

153. Fodor IK, Kamath C. Denoising through wavelet shrinkage: an empirical study. J Electron Imaging. 2003;12(1):151-60.

154. Vishwakarma DK, Kapoor R, Dhiman A, Goyal A, Jamil D. De-noising of Audio Signal Using Heavy Tailed Distribution and Comparison of Wavelets and Thresholding

Techniques. 2015 2nd International Conference on Computing for Sustainable Global Development (Indiacom). 2015:755-60.

155. Awal MA, Mostafa SS, Ahmad M, Rashid MA. An adaptive level dependent wavelet thresholding for ECG denoising. Biocybern Biomed Eng. 2014;34(4):238-49.

156. Ustundag M, Gokbulut M, Sengur A, Ata F. Denoising of weak ECG signals by using wavelet analysis and fuzzy thresholding. Netw Model Anal Hlth. 2012;1(4):135-40.

157. Rao RM, Slamani MA, Chyba TH, Emge DK. Wavelet-based denoising and baseline correction for enhancing chemical detection. Proc Spie. 2010;7698.

158. Zhang B, Sun LX, Yu HB, Xin Y, Cong ZB. A method for improving wavelet threshold denoising in laser-induced breakdown spectroscopy. Spectrochim Acta B. 2015;107:32-44.

159. Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, Hung MC, Kuerer HM. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. Proteomics. 2005;5(16):4107-17.

160. Abry P, Gonçalvés P, Flandrin P. Wavelets, spectrum analysis and 1/f processes. Wavelets and statistics: Springer; 1995. p. 15-29.

161. Wornell GW. Wavelet-Based Representations for the 1/F Family of Fractal Processes. P Ieee. 1993;81(10):1428-50.

162. Lee DTL, Yamamoto A. Wavelet Analysis - Theory and Applications. Hewlett-Packard J. 1994;45(6):44-54.

163. Gröchenig K. Foundations of time-frequency analysis : with 15 figures. Boston: Birkhäuser; 2001. xv, 359 p. p.

164. Allen JB. Short-Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier-Transform. Ieee T Acoust Speech. 1977;25(3):235-8.

165. Sejdic E, Djurovic I, Jiang J. Time-frequency feature representation using energy concentration: An overview of recent advances. Digit Signal Process. 2009;19(1):153-83.

166. Shensa MJ. The Discrete Wavelet Transform - Wedding the a Trous and Mallat Algorithms. Ieee T Signal Proces. 1992;40(10):2464-82.

167. Starck JL, Fadili J, Murtagh F. The undecimated wavelet decomposition and its reconstruction. Ieee T Image Process. 2007;16(2):297-309.

168. Beylkin G, Coifman R, Rokhlin V. Fast wavelet transforms and numerical algorithms I. Communications on pure and applied mathematics. 1991;44(2):141-83.

169. Daubechies I. Ten lectures on wavelets: Siam; 1992.

170. Jiang JA, Chao CF, Chiu MJ, Lee RG, Tseng CL, Lin R. An automatic analysis method for detecting and eliminating ECG artifacts in EEG. Comput Biol Med. 2007;37(11):1660-71.

171. Nenadic Z, Burdick JW. Spike detection using the continuous wavelet transform. Ieee T Bio-Med Eng. 2005;52(1):74-87.

172. Gisiger T. Scale invariance in biology: coincidence or footprint of a universal mechanism? Biol Rev. 2001;76(2):161-209.

173. Handel P, Chung A. Noise in physical systems and 1/f fluctuations. 1993.

174. Weissman MB. 1/F Noise and Other Slow, Nonexponential Kinetics in Condensed Matter. Rev Mod Phys. 1988;60(2):537-71.

175. Schmid H. Offset, flicker noise, and ways to deal with them. Circuits at the Nanoscale: CRC Press; 2018. p. 95-115.

176. Donoho DL, Johnstone IM. Adapting to unknown smoothness via wavelet shrinkage. J Am Stat Assoc. 1995;90(432):1200-24.

177. Xi J, Chen J. Noise Reduction Comparison Based on Different Wavelet Bases and Thresholds. Future Communication, Computing, Control and Management: Springer; 2012. p. 473-9.

178. Donoho DL. De-Noising by Soft-Thresholding. Ieee T Inform Theory. 1995;41(3):613-27.

179. Gao HY. Wavelet shrinkage denoising using the non-negative garrote. J Comput Graph Stat. 1998;7(4):469-88.

180. Agarwal S, Singh O, Nagaria D. Analysis and comparison of wavelet transforms for denoising MRI image. Biomedical and Pharmacology Journal. 2017;10(2):831-6.

181. Shao L, Yan RM, Li XL, Liu Y. From Heuristic Optimization to Dictionary Learning: A Review and Comprehensive Comparison of Image Denoising Algorithms. Ieee T Cybernetics. 2014;44(7):1001-13.

182. Costescu BI, Sturm S, Grater F. Dynamic disorder can explain non-exponential kinetics of fast protein mechanical unfolding. J Struct Biol. 2017;197(1):43-9.

183. Condat L. A Direct Algorithm for 1-D Total Variation Denoising. Ieee Signal Proc Let. 2013;20(11):1054-7.

184. Lu GH, Brittain JS, Holland P, Yianni J, Green AL, Stein JF, et al. Removing ECG noise from surface EMG signals using adaptive filtering. Neurosci Lett. 2009;462(1):14-9.

185. Hwang H, Haddad RA. Adaptive Median Filters - New Algorithms and Results. Ieee T Image Process. 1995;4(4):499-502.

186. Sharma T, Sharma KK. QRS complex detection in ECG signals using locally adaptive weighted total variation denoising. Comput Biol Med. 2017;87:187-99.

187. Zhou WF, Li QG. Adaptive total variation regularization based scheme for Poisson noise removal. Math Method Appl Sci. 2013;36(3):290-9.

188. Bronson JE, Fei JY, Hofman JM, Gonzalez RL, Wiggins CH. Learning Rates and States from Biophysical Time Series: A Bayesian Approach to Model Selection and Single-Molecule FRET Data. Biophys J. 2009;97(12):3196-205.

189. McKinney SA, Joo C, Ha T. Analysis of single-molecule FRET trajectories using hidden Markov modeling. Biophys J. 2006;91(5):1941-51.

190. Shuang B, Cooper D, Taylor JN, Kisley L, Chen JX, Wang WX, et al. Fast Step Transition and State Identification (STaSI) for Discrete Single-Molecule Data Analysis. J Phys Chem Lett. 2014;5(18):3157-61.

191. Dhindsa K. Filter-Bank Artifact Rejection: High performance real-time single-channel artifact detection for EEG. Biomed Signal Proces. 2017;38:224-35.

192. Islam MK, Rastegarnia A, Yang Z. Methods for artifact detection and removal from scalp EEG: A review. Neurophysiol Clin. 2016;46(4-5):287-305.

193. Junghofer M, Elbert T, Tucker DM, Rockstroh B. Statistical control of artifacts in dense array EEG/MEG studies. Psychophysiology. 2000;37(4):523-32.

194. de Cheveigne A. Time-shift denoising source separation. J Neurosci Meth. 2010;189(1):113-20.

195. Nolan H, Whelan R, Reilly RB. FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. J Neurosci Meth. 2010;192(1):152-62.

196. Ledergerber C, Dessimoz C. Base-calling for next-generation sequencing platforms. Brief Bioinform. 2011;12(5):489-97.

197. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12(6):443-51.

198. Choi Y, Moody IS, Sims PC, Hunt SR, Corso BL, Seitz DE, et al. Single-Molecule Dynamics of Lysozyme Processing Distinguishes Linear and Cross-Linked Peptidoglycan Substrates. Journal of the American Chemical Society. 2012;134(4):2032-5.

199. Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York: Wiley; 2001. xx, 654 p. p.

200. Landry M, Winters-Hilt S. Analysis of nanopore detector measurements using Machine-Learning methods, with application to single-molecule kinetic analysis. Bmc Bioinformatics. 2007;8.

201. Winters-Hilt S. Nanopore Detector based analysis of single-molecule conformational kinetics and binding interactions. Bmc Bioinformatics. 2006;7.

202. Albrecht T, Slabaugh G, Alonso E, Al-Arif SMMR. Deep learning for single-molecule science. Nanotechnology. 2017;28(42).

203. Thornley D, Petridis S. Machine learning in basecalling - Decoding trace peak behaviour. Proceedings of the 2006 Ieee Symposium on Computational Intelligence in Bioinformatics and Computational Biology. 2006:209-+.

204. Kircher M, Stenzel U, Kelso J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol. 2009;10(8).

205. Massingham T, Goldman N. All Your Base: a fast and accurate probabilistic approach to base calling. Genome Biol. 2012;13(2).

206. van den Akker J, Mishne G, Zimmer AD, Zhou AY. A machine learning model to determine the accuracy of variant calls in capture-based next generation sequencing. Bmc Genomics. 2018;19.

207. Luo RB, Sedlazeck FJ, Lam TW, Schatz MC. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat Commun. 2019;10.

208. Restrepo-Perez L, Joo C, Dekker C. Paving the way to single-molecule protein sequencing. Nat Nanotechnol. 2018;13(9):786-96.

209. Fleming MK, Cottrell GW. Categorization of Faces Using Unsupervised Feature-Extraction. Ieee Ijcnn. 1990:B65-B70.

210. Brunelli R, Poggio T. Face Recognition - Features Versus Templates. Ieee T Pattern Anal. 1993;15(10):1042-52.

211. Mahmoodabadi SZ, Ahmadian A, Abolhasani MD, Eslami M, Bidgoli JH. ECG feature extraction based on multiresolution wavelet transform. P Ann Int Ieee Embs. 2005:3902-5.

212. Subasi A. EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Syst Appl. 2007;32(4):1084-93.

213. Jenke R, Peer A, Buss M. Feature Extraction and Selection for Emotion Recognition from EEG. Ieee T Affect Comput. 2014;5(3):327-39.

214. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995;20(3):273-97.

215. Pearson K. On lines and planes of closest fit to systems of points in space. Philos Mag. 1901;2(7-12):559-72.

216. Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer; 2002. xxix, 487 p. p.

217. Scholkopf B, Smola A, Muller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 1998;10(5):1299-319.

218. Johnstone IM, Lu AY. On Consistency and Sparsity for Principal Components Analysis in High Dimensions Rejoinder. J Am Stat Assoc. 2009;104(486):701-3.

219. Collins M, Dasgupta S, Schapire RE. A generalization of principal component analysis to the exponential family. Adv Neur In. 2002;14:617-24.

220. Li J, Tao DC. Simple Exponential Family PCA. Ieee T Neur Net Lear. 2013;24(3):485-97.

221. Lu M, Huang JHZ, Qian XN. Sparse exponential family Principal Component Analysis. Pattern Recogn. 2016;60:681-91.

222. Jolliffe IT. A Note on the Use of Principal Components in Regression. Appl Stat-J Roy St C. 1982;31(3):300-3.

223. Kramer MA. Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. Aiche J. 1991;37(2):233-43.

224. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. J Mach Learn Res. 2010;11:3371-408.

225. Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nature Methods. 2008;5(8):679-82.

APPENDIX A

Lists of Features

A.1 Full List of Features

The full list of features, as used in the analyses in Sections 5.2.2 through 5.6, are given in Table A.1. An explanation of the various categories of features is given in Table 5.1.

#	Feature Name	#	Feature Name
Event Timing			
1	Event Duration		
Raw $I(t)$ / Denoised $I(t)$ within an event			
2	Raw I(t) Current Max (A)	17	Denoised I(t) Current Max (A)
3	Raw I(t) Current Min (A)	18	Denoised I(t) Current Min (A)
4	Raw I(t) Current Mean (A)	19	Denoised I(t) Current Mean (A)
5	Raw I(t) Current Std. Dev. (A)	20	Denoised I(t) Current Std. Dev. (A)
6	Raw I(t) Event Amplitude by Mean (A)	21	Denoised I(t) Event Amplitude by Mean (A)
7	Raw I(t) Current Median (A)	22	Denoised I(t) Current Median (A)
8	Raw I(t) Current Lower Quartile (A)	23	Denoised I(t) Current Lower Quartile (A)
9	Raw I(t) Current Upper Quartile (A)	24	Denoised I(t) Current Upper Quartile (A)
10	Raw I(t) Current Med. Abs. Dev. (A)	25	Denoised I(t) Current Med. Abs. Dev. (A)
11	Raw I(t) Event Amplitude by Median (A)	26	Denoised I(t) Event Amplitude by Median (A)
12	Raw I(t) Current Total Variation (A)	27	Denoised I(t) Current Total Variation (A)
13	Raw I(t) Current Total Variation (A) Per Unit Time	28	Denoised I(t) Current Total Variation (A) Per Unit Time
14	Raw I(t) Current Skew	29	Denoised I(t) Current Skew
15	Raw I(t) Current Kurtosis	30	Denoised I(t) Current Kurtosis
16	Raw I(t) Current Square Root of Sum of FFT Power (A)	31	Denoised I(t) Current Square Root of Sum of FFT Power (A)
UWT Scale Mean + Standard Deviation			
32	UWT Scale 10 Mean	41	UWT Scale 6 Std. Dev.
33	UWT Scale 10 Std. Dev.	42	UWT Scale 5 Mean
34	UWT Scale 9 Mean	43	UWT Scale 5 Std. Dev.
35	UWT Scale 9 Std. Dev.	44	UWT Scale 4 Mean
36	UWT Scale 8 Mean	45	UWT Scale 4 Std. Dev.
37	UWT Scale 8 Std. Dev.	46	UWT Scale 3 Mean
38	UWT Scale 7 Mean	47	UWT Scale 3 Std. Dev.
39	UWT Scale 7 Std. Dev.	48	UWT Scale 2 Mean
40	UWT Scale 6 Mean	49	UWT Scale 2 Std. Dev.

UWT Scale Median + MAD			
50	UWT Scale 10 Median	59	UWT Scale 6 Med. Abs. Dev.
51	UWT Scale 10 Med. Abs. Dev.	60	UWT Scale 5 Median
52	UWT Scale 9 Median	61	UWT Scale 5 Med. Abs. Dev.
53	UWT Scale 9 Med. Abs. Dev.	62	UWT Scale 4 Median
54	UWT Scale 8 Median	63	UWT Scale 4 Med. Abs. Dev.
55	UWT Scale 8 Med. Abs. Dev.	64	UWT Scale 3 Median
56	UWT Scale 7 Median	65	UWT Scale 3 Med. Abs. Dev.
57	UWT Scale 7 Med. Abs. Dev.	66	UWT Scale 2 Median
58	UWT Scale 6 Median	67	UWT Scale 2 Med. Abs. Dev.
UWT Scale Total Variation + Total Variation Per Unit Time			
68	UWT Scale 10 Total Variation	77	UWT Scale 6 Total Variation Per Unit Time
69	UWT Scale 10 Total Variation Per Unit Time	78	UWT Scale 5 Total Variation
70	UWT Scale 9 Total Variation	79	UWT Scale 5 Total Variation Per Unit Time
71	UWT Scale 9 Total Variation Per Unit Time	80	UWT Scale 4 Total Variation
72	UWT Scale 8 Total Variation	81	UWT Scale 4 Total Variation Per Unit Time
73	UWT Scale 8 Total Variation Per Unit Time	82	UWT Scale 3 Total Variation
74	UWT Scale 7 Total Variation	83	UWT Scale 3 Total Variation Per Unit Time
75	UWT Scale 7 Total Variation Per Unit Time	84	UWT Scale 2 Total Variation
76	UWT Scale 6 Total Variation	85	UWT Scale 2 Total Variation Per Unit Time
UWT Scale Skewness + Kurtosis			
86	UWT Scale 10 Skewness	95	UWT Scale 6 Kurtosis
87	UWT Scale 10 Kurtosis	96	UWT Scale 5 Skewness
88	UWT Scale 9 Skewness	97	UWT Scale 5 Kurtosis
89	UWT Scale 9 Kurtosis	98	UWT Scale 4 Skewness
90	UWT Scale 8 Skewness	99	UWT Scale 4 Kurtosis
91	UWT Scale 8 Kurtosis	100	UWT Scale 3 Skewness
92	UWT Scale 7 Skewness	101	UWT Scale 3 Kurtosis
93	UWT Scale 7 Kurtosis	102	UWT Scale 2 Skewness
94	UWT Scale 6 Skewness	103	UWT Scale 2 Kurtosis
UWT Scale Number of Zeros + Number of Zeros Per Unit Time			
104	UWT Scale 10 # of Zeros	113	UWT Scale 6 # of Zeros Per Unit Time
105	UWT Scale 10 # of Zeros Per Unit Time	114	UWT Scale 5 # of Zeros
106	UWT Scale 9 # of Zeros	115	UWT Scale 5 # of Zeros Per Unit Time
107	UWT Scale 9 # of Zeros Per Unit Time	116	UWT Scale 4 # of Zeros
108	UWT Scale 8 # of Zeros	117	UWT Scale 4 # of Zeros Per Unit Time
109	UWT Scale 8 # of Zeros Per Unit Time	118	UWT Scale 3 # of Zeros
110	UWT Scale 7 # of Zeros	119	UWT Scale 3 # of Zeros Per Unit Time
111	UWT Scale 7 # of Zeros Per Unit Time	120	UWT Scale 2 # of Zeros
112	UWT Scale 6 # of Zeros	121	UWT Scale 2 # of Zeros Per Unit Time
UWT Scale: Amplitude of Top 3 Peaks + Valleys			
122	UWT Scale 10 Peak 1 Amplitude	149	UWT Scale 6 Valley 1 Amplitude
123	UWT Scale 10 Peak 2 Amplitude	150	UWT Scale 6 Valley 2 Amplitude
124	UWT Scale 10 Peak 3 Amplitude	151	UWT Scale 6 Valley 3 Amplitude
125	UWT Scale 10 Valley 1 Amplitude	152	UWT Scale 5 Peak 1 Amplitude
126	UWT Scale 10 Valley 2 Amplitude	153	UWT Scale 5 Peak 2 Amplitude
127	UWT Scale 10 Valley 3 Amplitude	154	UWT Scale 5 Peak 3 Amplitude
128	UWT Scale 9 Peak 1 Amplitude	155	UWT Scale 5 Valley 1 Amplitude
129	UWT Scale 9 Peak 2 Amplitude	156	UWT Scale 5 Valley 2 Amplitude
130	UWT Scale 9 Peak 3 Amplitude	157	UWT Scale 5 Valley 3 Amplitude
131	UWT Scale 9 Valley 1 Amplitude	158	UWT Scale 4 Peak 1 Amplitude

132	UWT Scale 9 Valley 2 Amplitude	159	UWT Scale 4 Peak 2 Amplitude
133	UWT Scale 9 Valley 3 Amplitude	160	UWT Scale 4 Peak 3 Amplitude
134	UWT Scale 8 Peak 1 Amplitude	161	UWT Scale 4 Valley 1 Amplitude
135	UWT Scale 8 Peak 2 Amplitude	162	UWT Scale 4 Valley 2 Amplitude
136	UWT Scale 8 Peak 3 Amplitude	163	UWT Scale 4 Valley 3 Amplitude
137	UWT Scale 8 Valley 1 Amplitude	164	UWT Scale 3 Peak 1 Amplitude
138	UWT Scale 8 Valley 2 Amplitude	165	UWT Scale 3 Peak 2 Amplitude
139	UWT Scale 8 Valley 3 Amplitude	166	UWT Scale 3 Peak 3 Amplitude
140	UWT Scale 7 Peak 1 Amplitude	167	UWT Scale 3 Valley 1 Amplitude
141	UWT Scale 7 Peak 2 Amplitude	168	UWT Scale 3 Valley 2 Amplitude
142	UWT Scale 7 Peak 3 Amplitude	169	UWT Scale 3 Valley 3 Amplitude
143	UWT Scale 7 Valley 1 Amplitude	170	UWT Scale 2 Peak 1 Amplitude
144	UWT Scale 7 Valley 2 Amplitude	171	UWT Scale 2 Peak 2 Amplitude
145	UWT Scale 7 Valley 3 Amplitude	172	UWT Scale 2 Peak 3 Amplitude
146	UWT Scale 6 Peak 1 Amplitude	173	UWT Scale 2 Valley 1 Amplitude
147	UWT Scale 6 Peak 2 Amplitude	174	UWT Scale 2 Valley 2 Amplitude
148	UWT Scale 6 Peak 3 Amplitude	175	UWT Scale 2 Valley 3 Amplitude
UWT Product of Two Scales			
176	UWT Scale 5*6 Product - # of Significant Peaks	182	UWT Scale 4*5 Product - Peak 2 Amplitude
177	UWT Scale 5*6 Product - Peak 1 Amplitude	183	UWT Scale 4*5 Product - Peak 3 Amplitude
178	UWT Scale 5*6 Product - Peak 2 Amplitude	184	UWT Scale 3*4 Product - # of Significant Peaks
179	UWT Scale 5*6 Product - Peak 3 Amplitude	185	UWT Scale 3*4 Product - Peak 1 Amplitude
180	UWT Scale 4*5 Product - # of Significant Peaks	186	UWT Scale 3*4 Product - Peak 2 Amplitude
181	UWT Scale 4*5 Product - Peak 1 Amplitude	187	UWT Scale 3*4 Product - Peak 3 Amplitude
Features From Multiple Channels			
188	# of MidPass Events Per LowPass Event	189	# of HighPass Events Per LowPass Event

Table A.1: Full list of the feature numbers and names included in the analysis discussed in Chapter 5. A description of the feature categories is contained in Table 5.1.

A.2 Composition of Principal Components

Principal components are eigenvectors which group correlated or similar features together and map the grouped features into the same dimension in a new, usually reduced-size, vector space. By definition, if $v_{i,n}$ are the individual coefficients of the *n*th principal component $\mathbf{v}_n = v_{i,n}\hat{\mathbf{u}}_i$, where the $\hat{\mathbf{u}}_i$ represent the individual features in the original feature space, the sum of the squares of $v_{i,n}$ is 1: $\sum_i v_{i,n}^2 = 1$. The magnitudes of the individual coordinates indicate the relative contribution of the corresponding feature to

the principal component. The following five tables (Tables A.2 – A.6) display the names of the most significant features in the first 4 principal components, defined by having a coefficient magnitude $|v_{i,n}| \geq 0.1$. Within each table, the features are listed from largest to smallest magnitude.

Principal Component Number	Principal Component Description		
1	UWT Peak Heights		
Feature Name	Coefficient Value	Feature Name	Coefficient Value
UWT Scale 4*5 Product - Peak 2 Amplitude	0.131558	UWT Scale 7 Valley 1 Amplitude	0.120932
UWT Scale 5 Peak 2 Amplitude	0.131308	Raw I(t) Current Square Root of Sum of FFT Power (A)	0.11943
UWT Scale 6 Peak 2 Amplitude	0.131175	UWT Scale 6 Valley 1 Amplitude	0.119154
UWT Scale 4 Peak 2 Amplitude	0.130565	Denoised I(t) Current Total Variation (A)	0.116577
UWT Scale 3*4 Product - Peak 3 Amplitude	0.130549	UWT Scale 6 Valley 2 Amplitude	0.115969
UWT Scale 3*4 Product - Peak 2 Amplitude	0.130155	UWT Scale 3 Peak 1 Amplitude	0.115873
UWT Scale 4*5 Product - Peak 3 Amplitude	0.12921	UWT Scale 5 Valley 3 Amplitude	0.115079
UWT Scale 6 Peak 1 Amplitude	0.128305	UWT Scale 7 Valley 3 Amplitude	0.114608
UWT Scale 5 Peak 1 Amplitude	0.127478	UWT Scale 6 Valley 3 Amplitude	0.114448
UWT Scale 4*5 Product - Peak 1 Amplitude	0.126883	UWT Scale 5 Valley 2 Amplitude	0.112739
UWT Scale 3*4 Product - Peak 1 Amplitude	0.125048	UWT Scale 5 Valley 1 Amplitude	0.112402
UWT Scale 3 Peak 2 Amplitude	0.124931	UWT Scale 2 Peak 2 Amplitude	0.108766
UWT Scale 5*6 Product - Peak 2 Amplitude	0.123888	Raw I(t) Current Med. Abs. Dev. (A)	0.10713
UWT Scale 7 Peak 2 Amplitude	0.122699	Denoised I(t) Current Std. Dev. (A)	0.103314
UWT Scale 4 Peak 1 Amplitude	0.122141	UWT Scale 8 Peak 3 Amplitude	0.101813
UWT Scale 5*6 Product - Peak 3 Amplitude	0.121968	Denoised I(t) Current Max (A)	0.101658
UWT Scale 7 Peak 3 Amplitude	0.121495	UWT Scale 3 Valley 2 Amplitude	0.101008
UWT Scale 5*6 Product - Peak 1 Amplitude	0.121462	UWT Scale 2 Peak 1 Amplitude	0.10055
Raw I(t) Current Std. Dev. (A)	0.121297		

Table A.2: List of the most-significant features, and the coefficients corresponding to each, in principal component 1, ranked from largest to smallest.

Principal Component Number	Principal Component Description		
2	Event Duration		
Feature Name	Coefficient Value	Feature Name	Coefficient Value
UWT Scale 8 # of Zeros	0.150758	UWT Scale 6 Total Variation	0.148025
UWT Scale 6 # of Zeros	0.150335	UWT Scale 7 Total Variation	0.146974
UWT Scale 7 # of Zeros	0.15028	UWT Scale 8 Total Variation	0.146218
UWT Scale 9 # of Zeros	0.150156	UWT Scale 9 Total Variation	0.145192
UWT Scale 4 # of Zeros	0.150146	UWT Scale 10 Total Variation	0.144721
UWT Scale 5 # of Zeros	0.149917	# of MidPass Events Per LowPass Event	0.139351
UWT Scale 2 # of Zeros	0.149797	UWT Scale 6 Std. Dev.	-0.118281
UWT Scale 2 Total Variation	0.14976	UWT Scale 5 Std. Dev.	-0.117548
Raw I(t) Current Total Variation (A)	0.149535	Denoised I(t) Current Total Variation (A)	0.112023
Event Duration (s)	0.149487	Denoised I(t) Current Total Variation (A) Per Unit Time	-0.111151
UWT Scale 3 # of Zeros	0.149465	UWT Scale 7 Std. Dev.	-0.110575
UWT Scale 4 Total Variation	0.149273	UWT Scale 4 Std. Dev.	-0.108667
UWT Scale 3*4 Product - # of Significant Peaks	0.149196	UWT Scale 7 Total Variation Per Unit Time	-0.107013
UWT Scale 10 # of Zeros	0.149081	UWT Scale 6 Total Variation Per Unit Time	-0.103456
UWT Scale 3 Total Variation	0.148837	Raw I(t) Current Square Root of Sum of FFT Power (A)	-0.102966
UWT Scale 4*5 Product - # of Significant Peaks	0.14877	Raw I(t) Current Std. Dev. (A)	-0.102918
UWT Scale 5*6 Product - # of Significant Peaks	0.148644	UWT Scale 5 Total Variation Per Unit Time	-0.101072
UWT Scale 5 Total Variation	0.148244		

Table A.3: List of the most-significant features, and the coefficients corresponding to each, in principal component 2, ranked from largest to smallest.

Principal Component Number	Principal Component Description		
3	Noise Strength		
Feature Name	Coefficient Value	Feature Name	Coefficient Value
UWT Scale 9 Total Variation Per Unit Time	-0.18099	UWT Scale 5*6 Product - # of Significant Peaks	-0.105263
UWT Scale 10 Total Variation Per Unit Time	-0.174129	UWT Scale 7 # of Zeros	-0.104007
UWT Scale 8 Std. Dev.	-0.169478	UWT Scale 6 # of Zeros	-0.103967
UWT Scale 8 Total Variation Per Unit Time	-0.166452	UWT Scale 9 # of Zeros	-0.103932
UWT Scale 8 Med. Abs. Dev.	-0.148934	UWT Scale 5 # of Zeros	-0.103919
UWT Scale 9 Std. Dev.	-0.148348	UWT Scale 4*5 Product - # of Significant Peaks	-0.103819
UWT Scale 7 Total Variation Per Unit Time	-0.144079	UWT Scale 9 Peak 1 Amplitude	-0.103624

UWT Scale 9 Med. Abs. Dev.	-0.144074	UWT Scale 9 Total Variation	-0.103527
Raw I(t) Current Max (A)	0.142206	UWT Scale 4 # of Zeros	-0.1034
Raw I(t) Current Median (A)	0.142191	UWT Scale 3 Valley 2 Amplitude	0.103126
UWT Scale 7 Std. Dev.	-0.139867	UWT Scale 3 # of Zeros	-0.102999
UWT Scale 2 Valley 2 Amplitude	0.128023	UWT Scale 3*4 Product - # of Significant Peaks	-0.102994
UWT Scale 8 Peak 1 Amplitude	-0.115879	UWT Scale 8 # of Zeros	-0.102859
# of MidPass Events Per LowPass Event	-0.114553	UWT Scale 2 # of Zeros	-0.102665
Raw I(t) Event Amplitude by Median (A)	0.113996	UWT Scale 10 Total Variation	-0.102511
UWT Scale 4 Valley 2 Amplitude	0.112193	Event Duration (s)	-0.102477
UWT Scale 2 # of Zeros Per Unit Time	0.111732	UWT Scale 8 Total Variation	-0.102038
Denoised I(t) Current Total Variation (A) Per Unit Time	-0.110932	UWT Scale 2 Total Variation	-0.1015
Denoised I(t) Current Median (A)	0.110375	Raw I(t) Current Total Variation (A)	-0.101367
Raw I(t) Event Amplitude by Mean (A)	0.109269	UWT Scale 7 Total Variation	-0.101149
Denoised I(t) Current Mean (A)	0.107877	UWT Scale 3 Total Variation	-0.101097
UWT Scale 2 Valley 1 Amplitude	0.10687	UWT Scale 6 Total Variation	-0.100648
UWT Scale 10 # of Zeros	-0.105489	UWT Scale 4 Total Variation	-0.100633
UWT Scale 7 Med. Abs. Dev.	-0.105332	UWT Scale 5 Total Variation	-0.100599

Table A.4: List of the most-significant features, and the coefficients corresponding to each, in principal component 3, ranked from largest to smallest.

Principal Component Number	Principal Component Description		
4	UWT Mid-Frequency Mean Values		
Feature Name	Coefficient Value	Feature Name	Coefficient Value
UWT Scale 7 Mean	-0.260789	UWT Scale 9 Median	-0.215678
UWT Scale 8 Mean	-0.256609	UWT Scale 7 Median	-0.210428
UWT Scale 6 Mean	-0.252414	UWT Scale 6 Median	-0.192073
UWT Scale 5 Mean	-0.241611	UWT Scale 2 Mean	-0.191332
UWT Scale 4 Mean	-0.230763	UWT Scale 5 Median	-0.176705
UWT Scale 8 Median	-0.227134	UWT Scale 10 Median	-0.174619
UWT Scale 9 Mean	-0.21918	UWT Scale 10 Mean	-0.167218
UWT Scale 3 Mean	-0.218384	UWT Scale 4 Median	-0.108644

Table A.5: List of the most-significant features, and the coefficients corresponding to each, in principal component 4, ranked from largest to smallest.

Principal Component Number	Principal Component Description		
5	Event Amplitudes + Noise Amplitudes		
Feature Name	Coefficient Value	Feature Name	Coefficient Value
Denoised I(t) Current Upper Quartile (A)	-0.210303	UWT Scale 2 Med. Abs. Dev.	-0.156829
Denoised I(t) Event Amplitude by Mean (A)	-0.210003	UWT Scale 6 Total Variation Per Unit Time	-0.148447
Denoised I(t) Current Mean (A)	-0.206856	Denoised I(t) Current Max (A)	-0.145286
UWT Scale 10 Std. Dev.	0.202231	UWT Scale 3 Std. Dev.	-0.139282
Denoised I(t) Event Amplitude by Median (A)	-0.197033	Denoised I(t) Current Total Variation (A) Per Unit Time	-0.138127
Denoised I(t) Current Median (A)	-0.193568	UWT Scale 7 Total Variation Per Unit Time	-0.127606
UWT Scale 2 Std. Dev.	-0.188033	Raw I(t) Current Median (A)	-0.125243
UWT Scale 10 Peak 1 Amplitude	0.187391	UWT Scale 3 Med. Abs. Dev.	-0.121698
Raw I(t) Event Amplitude by Mean (A)	-0.185658	UWT Scale 5 Total Variation Per Unit Time	-0.121517
Raw I(t) Current Total Variation (A) Per Unit Time	-0.178943	Raw I(t) Current Max (A)	-0.118516
Raw I(t) Event Amplitude by Median (A)	-0.177828	UWT Scale 3 Total Variation Per Unit Time	-0.109624
Denoised I(t) Current Min (A)	-0.174252	UWT Scale 8 Kurtosis	0.109256
UWT Scale 2 Total Variation Per Unit Time	-0.170244	UWT Scale 4 Std. Dev.	-0.108062
UWT Scale 4 Total Variation Per Unit Time	-0.167002	UWT Scale 4 Med. Abs. Dev.	-0.102996
UWT Scale 10 Med. Abs. Dev.	0.164783		

Table A.6: List of the most-significant features, and the coefficients corresponding to each, in principal component 5, ranked from largest to smallest.

APPENDIX B

LabVIEW Programs

B.1 Introduction

One minor objective of the analysis procedures described in Chapters 4 and 5 is to automate the denoising and event identification as much as possible to increase data processing throughput, both by reducing the need for time-consuming manual analysis and by efficiently utilizing multiple computers and cores. The eventual goal is to automatically calculate estimates of biomolecule activity from measurements obtained on one day and have the results ready by the next day, so that new measurements can incorporate quantitative information gathered from the previous day's measurements.

The analysis of SWCNT-FET signals, as implemented in LabVIEW, is split into two parts:

1) Denoising, event identification, and calculation of feature values
2) Feature visualization, evaluation of correlations, and PCA

The first part is contained in one LabVIEW virtual instrument (VI) with several subVIs, while the second part is contained in four separate VIs. The next two sections briefly describe the functions of each VI.

B.2 Denoising, Event Identification, and Feature Selection

The general framework for selecting measurements, choosing denoising and event identification parameters, loading the data from disk, and running the signal processing routines is contained in the AutoAnalysis2 VI, whose front panel is displayed in Figure B.1. The front panel is organized as follows, with each section highlighted by an enclosing box:

- Left: source and save directory locations (light blue)
- Center top: configuration settings for signal processing (dark blue)
- Bottom right: analysis queue details and status of current analysis (red)

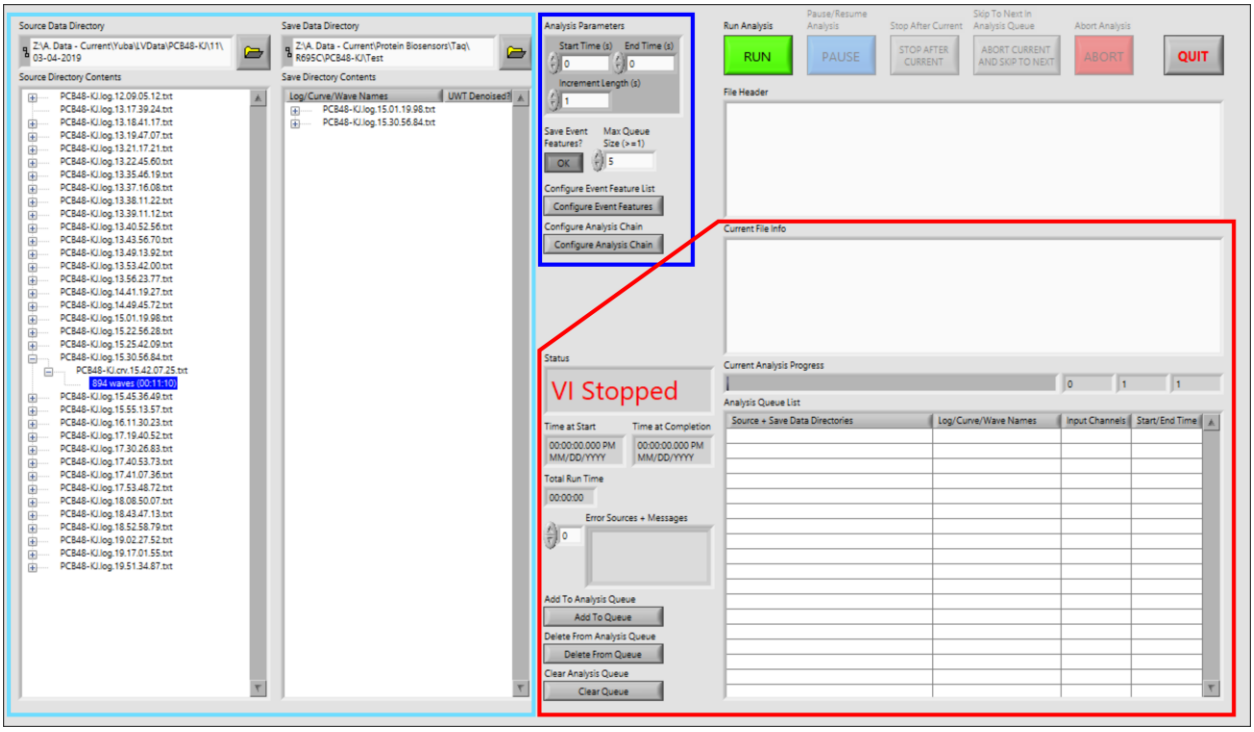

Figure B.1: Front panel of the "AutoAnalysis2" VI, which provides a framework for loading and processing multiple datasets.

The source directory is the location of the unprocessed data, and the save directory is where the processed signals and event information are saved. When an individual measurement from the source directory is selected, the measurement metadata is loaded and displayed in File Header (top right). The denoising and event identification parameters can be adjusted by selecting "Configure Analysis Chain", opening a window which allows for selection of various signal processing subVIs, as shown in Figure B.2. As implemented here, each signal processing subVI instance runs in its own memory space, allowing multiple copies of the subVIs to each maintain its own state. The wavelet 1/f denoising discussed in Chapter 4 is implemented as the "DenoiseUWT" subVI (Figure B.3), and event identification and feature value calculation is implemented as the "AutoMultiState_wStateVector" VI (Figure B.4), which also stores the event feature information in a spreadsheet. Calculation of feature categories can be turned on or off in the "Configure Event Features" window, shown in Figure B.5.

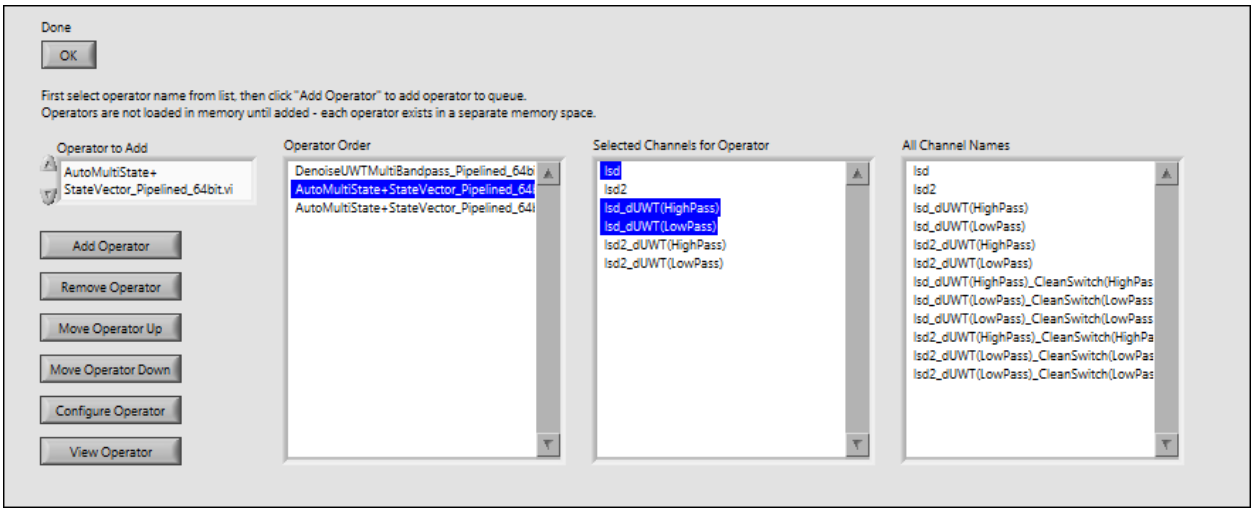

Figure B.2: Front panel of the "Configure Analysis Chains" subVI, which specifies the signal channels to process and the analysis subVIs to run.

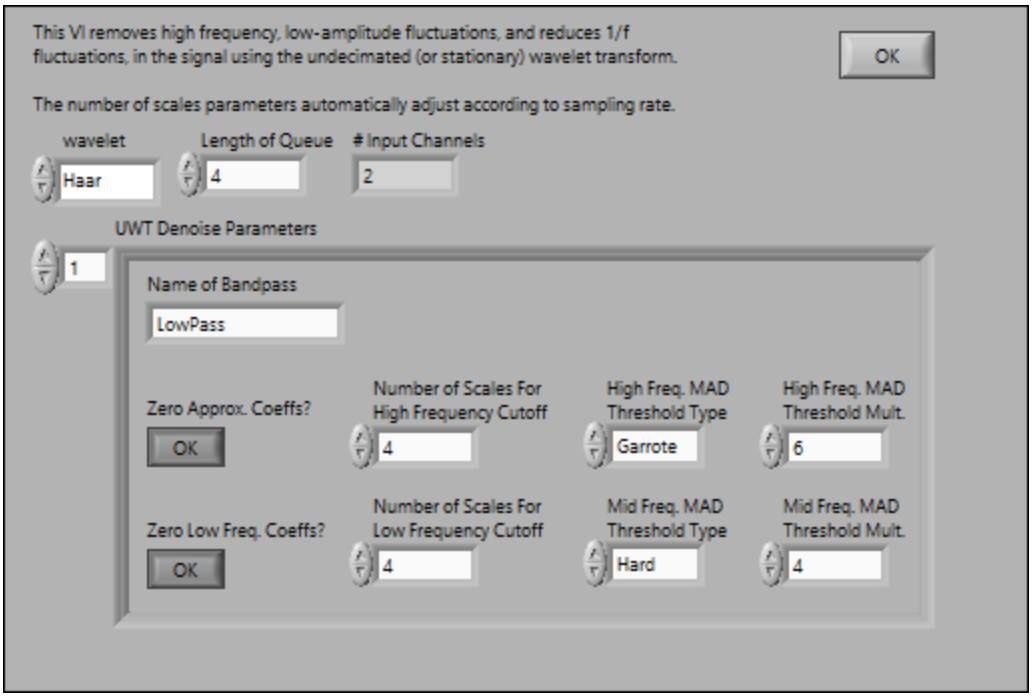

Figure B.3: Front panel of the "DenoiseUWTMultiBandpass" subVI, which performs the wavelet 1/f denoising as described in Chapter 4.

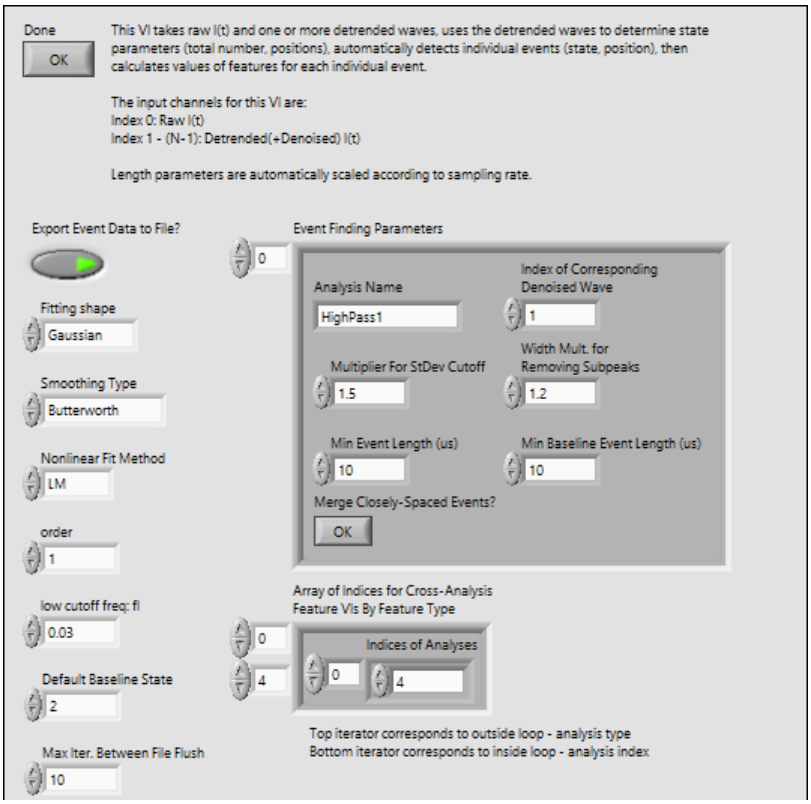

Figure B.4: Front panel of the "AutoMultiState_wStateVector" subVI, which performs automatic state identification and feature value calculation as described in Chapter 5.

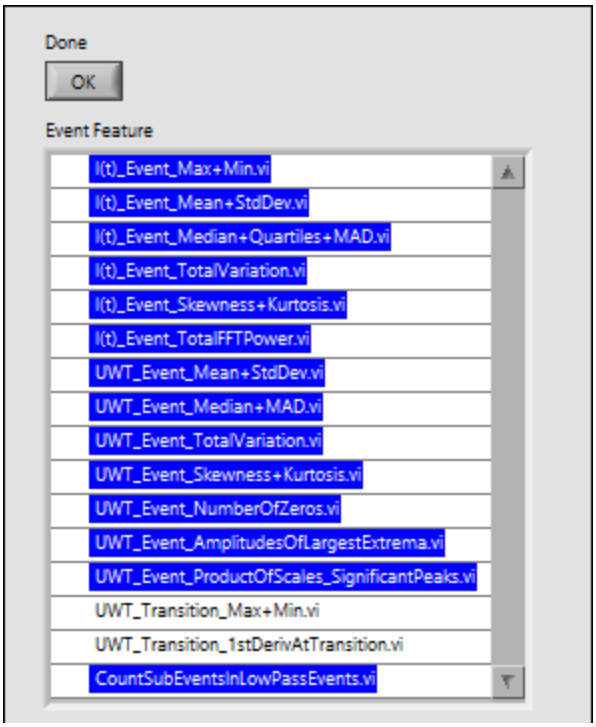

Figure B.5: Front panel of the "Configure Event Features" subVI, which specifies the feature categories that are calculated for each event.

Once the analysis chain has been configured, individual measurements can be added to the analysis queue, which stores a list of measurements and analysis information for sequential processing. Selecting "Run Analysis" pushes the first measurement from the queue into the execution loop, which is illustrated in the program schematic in Figure B.6. The data is first loaded from disk, then run through the signal processing pipeline in steps, and the results are written back to disk. Since the signal processing is performed in batches, each iteration of the loop operates on one batch, and the loops are designed to run in parallel. The pipelined execution gives the program flexibility to take advantage of multiple CPU threads and perform other signal processing steps without waiting for a previous step to complete.

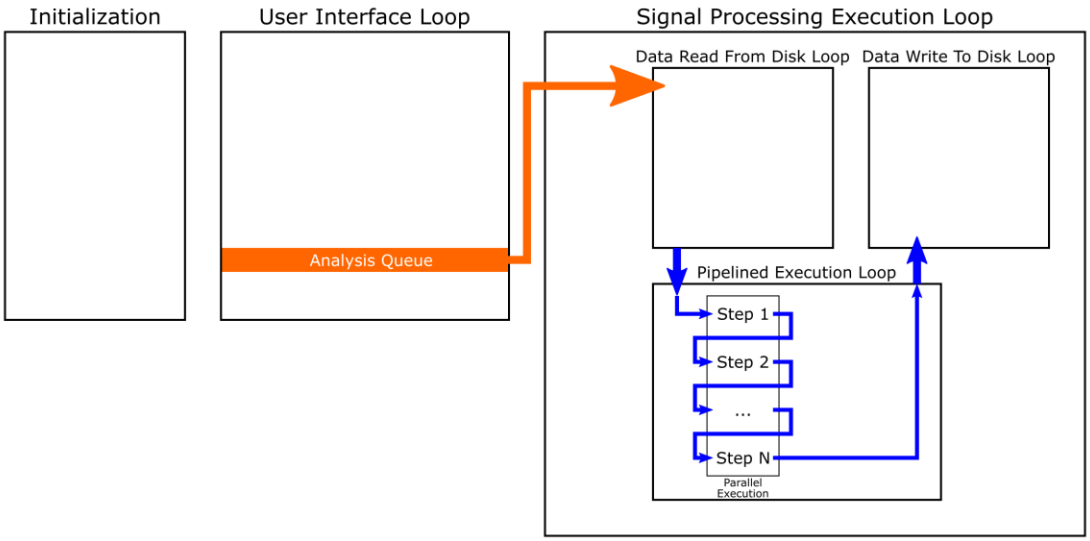

Figure B.6: Schematic of the "AutoAnalysis2" VI, showing the three different segments of the block diagram. The analysis queue and its information flow are shown in orange, while the data flow in the execution loop is shown in blue.

As a summary of the current state of analysis automation, the denoising procedures are robust and do not require parameter adjustments for each measurement, but the event identification procedure struggles with measurements exhibiting signal-to-noise ratios of less than 2:1 (as defined by the ratio between the event amplitude and the noise center-to-peak). Specifically, event identification selects many false positives, as many as 10 false positives to one true positive on low SNR measurements. This provides a sufficient starting point for manual selection of "clean" events through visual inspection, but more work is needed to optimize automatic event identification. One significant weakness in the algorithm is the instability of the state-identification procedure discussed in Section 5.2.1, which may fit too many Gaussian functions to the $I(t)$ histogram or place peak centers in incorrect locations.

B.3 Visualizing Individual Events and Feature Distributions

Once the event information is saved in the feature spreadsheet, the details can be viewed in several ways. Individual events (and the UWT coefficients corresponding to the event) can be viewed with the "DisplayIndividualEvents" VI (Figure B.7), which also provides a section for manually selecting events for later analysis. The "Event Number" control at the top of the VI facilitates scrolling between events of the same state, while the "Event Index" control indicates the row (in the 2D spreadsheet of event information) corresponding to that specific event. The event index values serve as the event identifiers when passing data between different VIs.

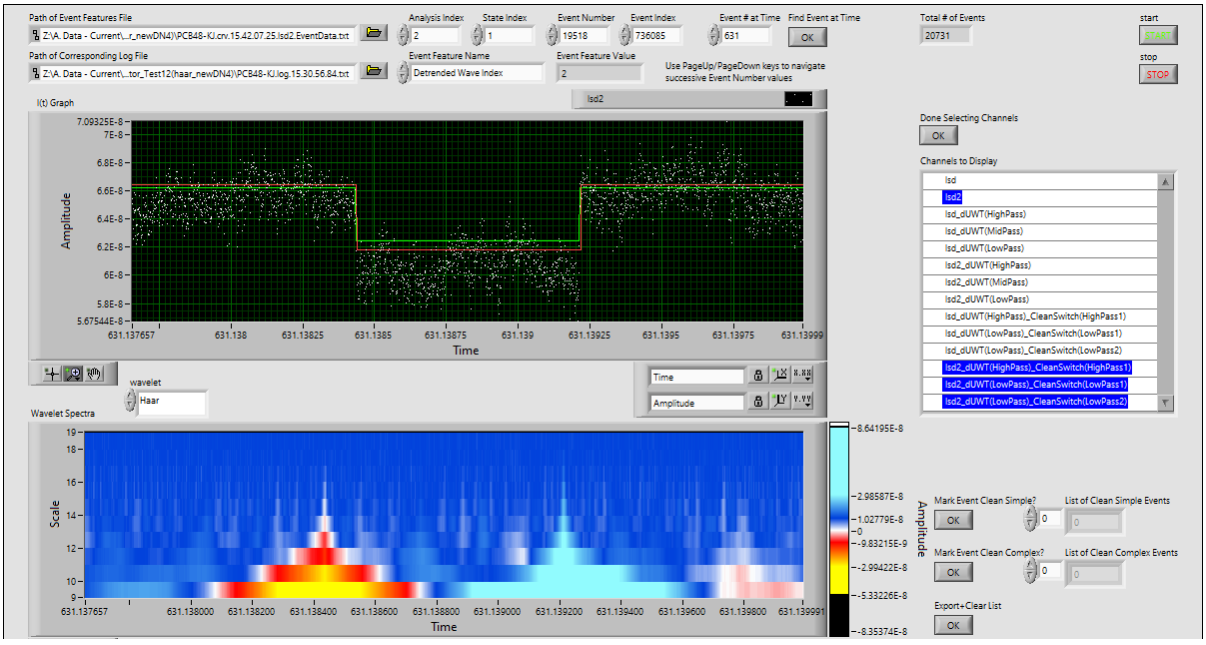

Figure B.7: Front panel of the "DisplayIndividualEvents" VI.

The “State+Duration_Histogram” VI (Figure B.8) displays the distribution of event durations (top graph) and the event rate over time (bottom graph).

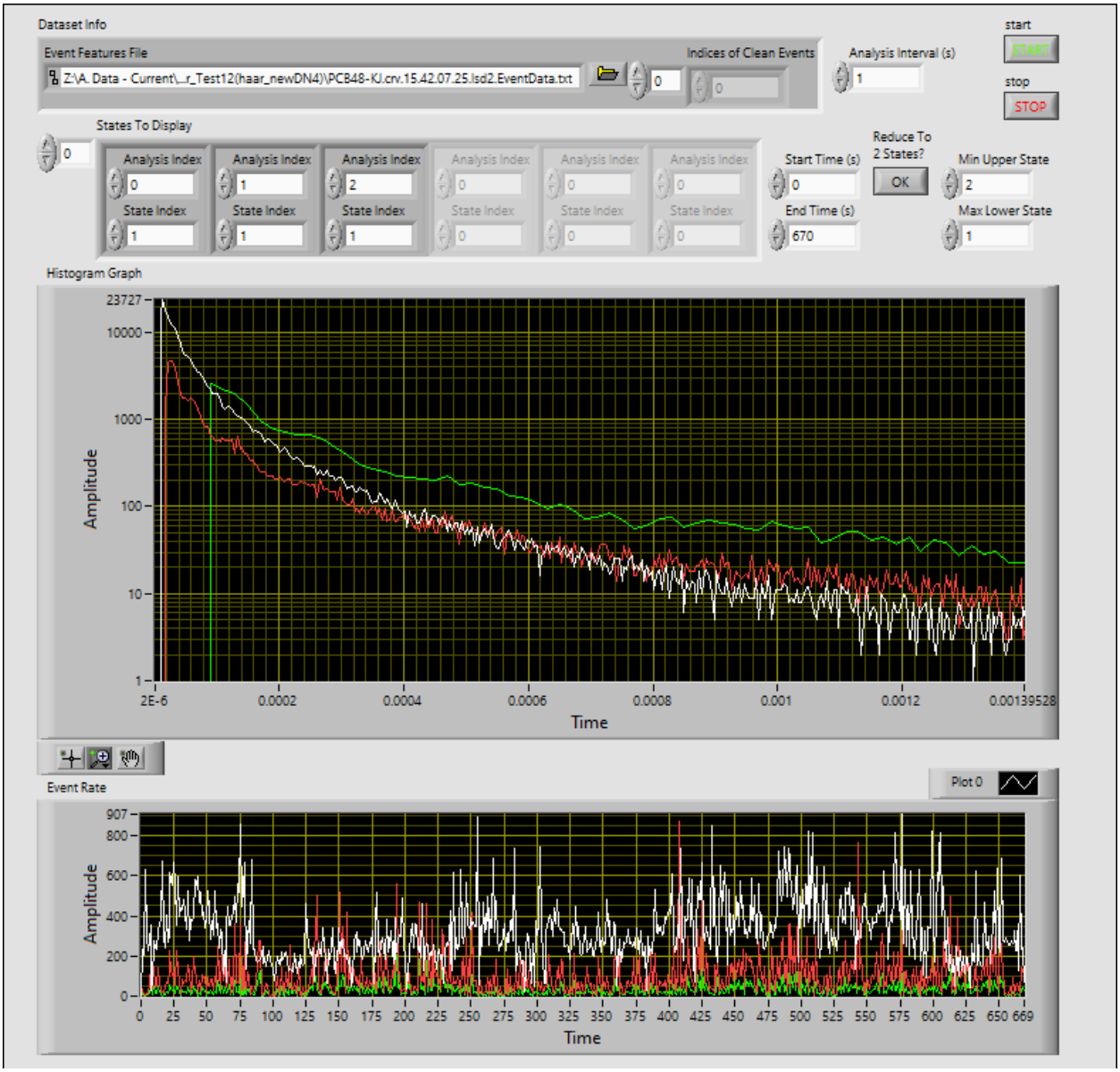

Figure B.8: Front panel of the “State+Duration_Histogram” VI.

Visual inspection of the feature distributions, correlations of distributions, and the principal components calculated from PCA are done in the “ComparisonXY” (Figure B.9)

and “ComparisonXYCleanEventsMultipleDatasets” (Figure B.10) VIs. Both VIs use a similar interface: two operating modes (original features vs. PCA components, toggled using the “Plot PCA Vectors” button), and two display types: “Points” to display the distribution of values from individual points, and “Correlation Matrix” to show the full correlation matrix. A picture of the “ComparisonXY” VI showing the correlation matrix is given in Figure B.11. In the “Points” display, the features displayed on the X and Y axes are changed with the “X Event Feature” and “Y Event Feature” controls, and the histograms underneath display the distributions of the data on each axis. In addition, the “Points” graph has a cursor (crossed yellow lines) which can identify the event index corresponding to a specific point, which can then be input into the “DisplayIndividualEvents” VI to view the $I(t)$ and UWT of that event.

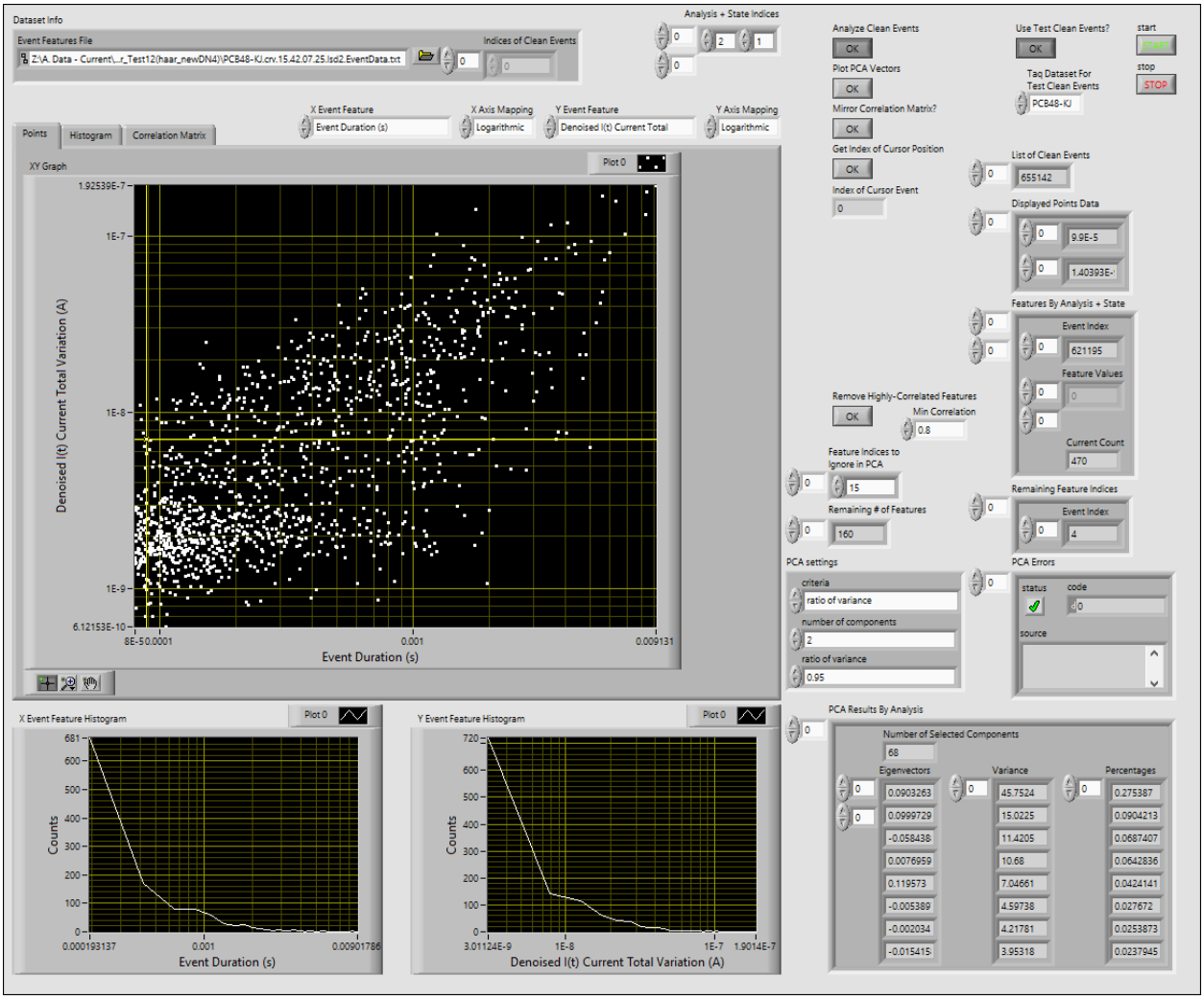

Figure B.9: Front panel of the "ComparisonXY" VI, set in the "PCA" mode and showing the "Points" display.

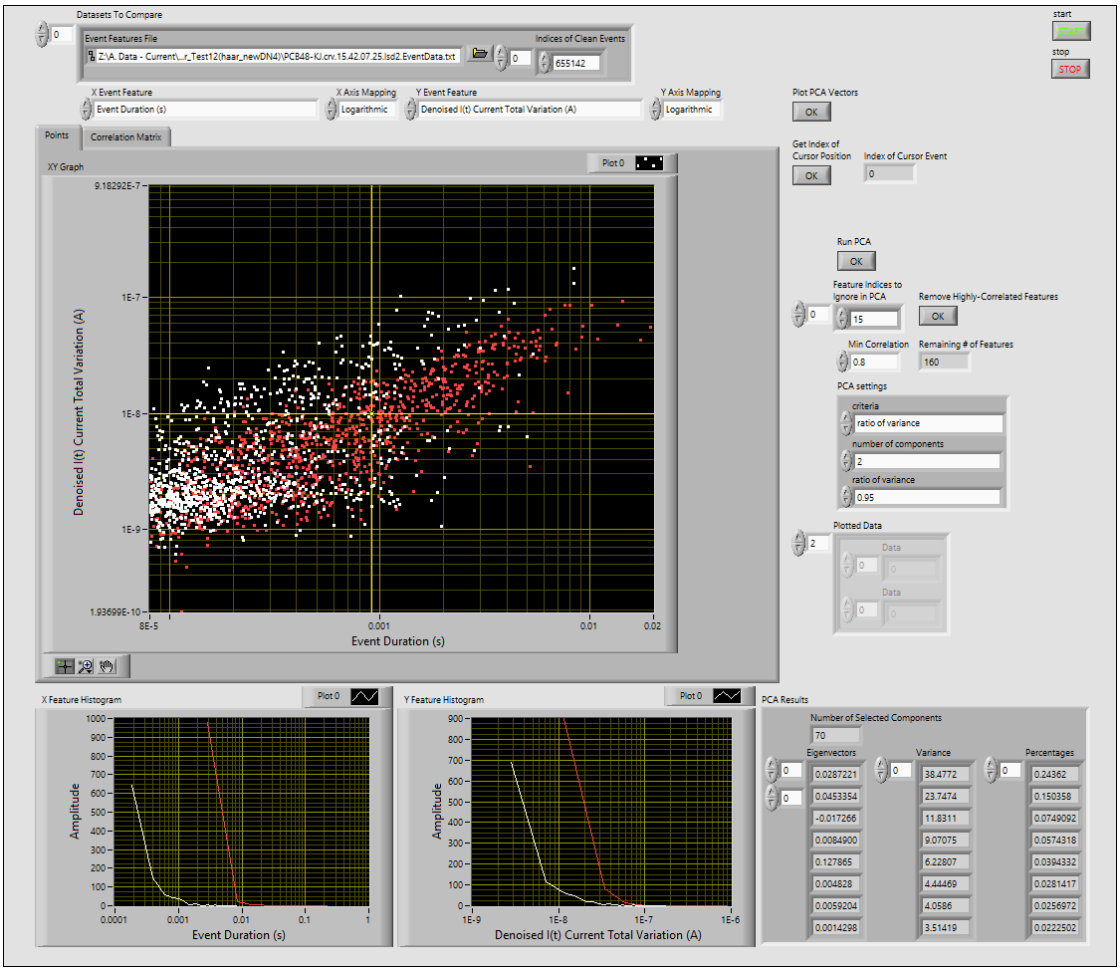

Figure B.10: Front panel of the "ComparisonXYCleanEventsMultipleDatasets" VI, set in the original data mode and showing the "Points" display. The data being displayed is from the PCB48-KJ and PCB48-JK datasets as discussed in Chapter 5.

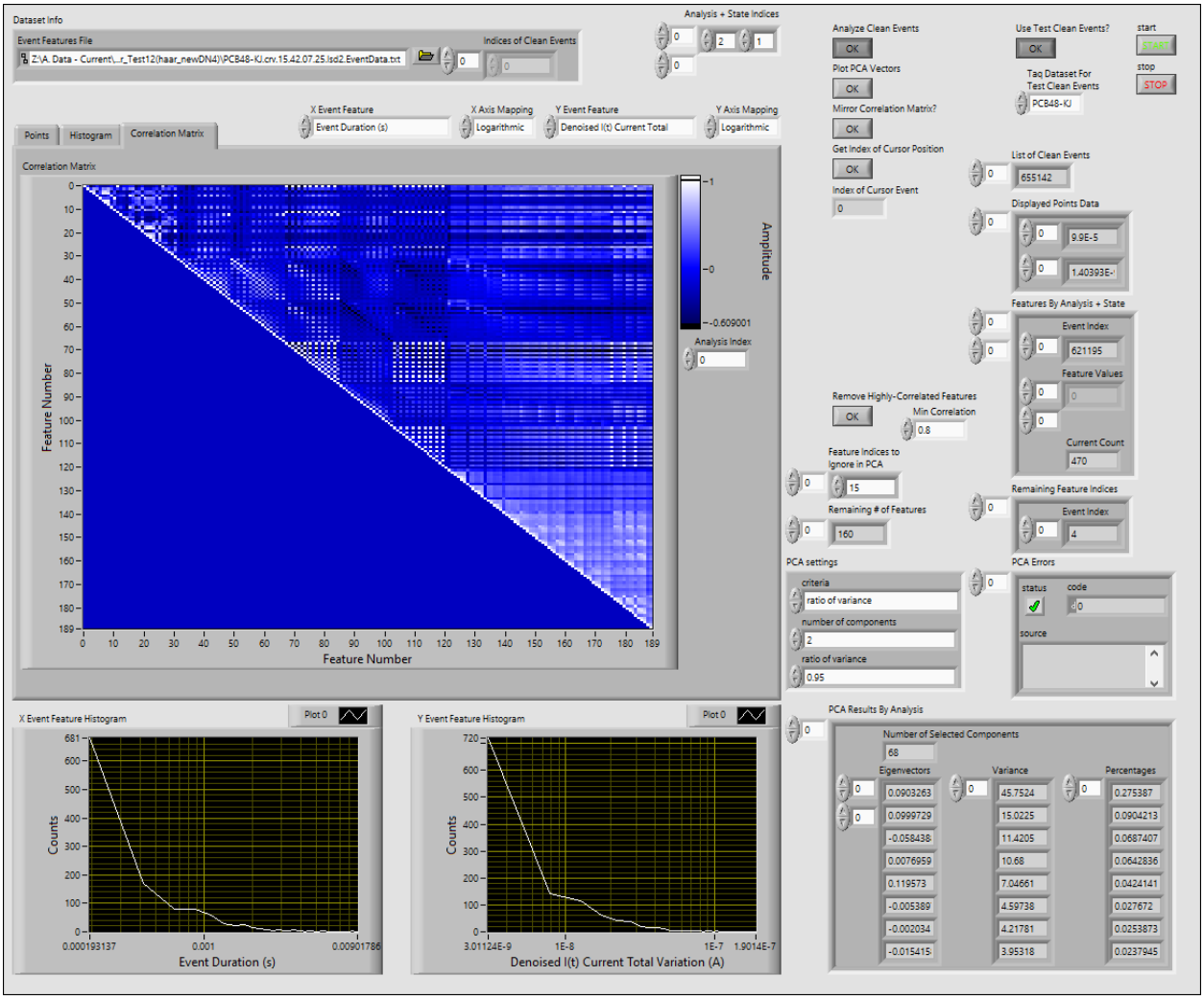

Figure B.11: Front panel of the "ComparisonXY" VI, set in the original data mode and showing the "Correlation Matrix" display.

Both the "ComparisonXY" and "ComparisonXYCleanEventsMultipleDatasets" VIs provide the option to view only manually-selected events, which input as an array of event indices as obtained from the "DisplayIndividualEvents" VI.