UNIVERSITY OF CALIFORNIA, SAN DIEGO

Simulation of a Scalable Electrochemical Immunosignaturing Biosensor Array for Disease Diagnosis

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Bioengineering

by

Anthony Au

Committee in charge:

Professor Vikash Gilja, Chair
Professor Gert Cauwenberghs, Co-Chair
Professor Stephanie Fraley
Professor Drew Hall

2015

The thesis of Anthony Au is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2015

DEDICATION

To my mom and dad for all of their support and encouragement throughout my life. Despite how it may have seemed at the time, they were trying their hardest to do what's best for me.

EPIGRAPH

You Have Brains In Your Head.

You Have Feet In Your Shoes.

You Can Steer Yourself

Any Direction You Choose.

—Dr. Seuss

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Firstly, I would like to acknowledge Professor Vikash Gilja for his support as the chair of my committee. His guidance has been an enormous help during the course of this work. I would have been lost at times without him steering me in the right direction.

Secondly, I would like to acknowledge Professor Drew Hall for his collaboration and expert advice on this project. I would also like to thank him for serving as part of my committee and giving me many constructive comments for further directions and improvements.

Thirdly, I would like to acknowledge Alexander Sun for answering my many questions and obtaining experimental data for comparison and validation of the simulations.

Fourthly, I would like to acknowledge A.G. Venkatesh for fabricating the physical sensors for preliminary experiments. It has proven to be a crucial element to my thesis.

Fifthly, I would like to acknowledge Professor Gert Cauwenberghs and Professor Stephanie Fraley for serving as part of my committee. Their unique perspectives and insights were helpful for determining future directions and emphasizing the importance of this work.

Last of all, I would like to acknowledge the members of TNEL, Yuchen Wang, Tejaswy Pailla, John Hermiz, Werner Jiang, Paolo Gabriel, Nick Rodgers, Francis Baek, Akinyinka Omigbodun, and Hannah Chou. Their support and feedback has been inestimable. Even for earlier projects, they would always be available for discussions and offer advice for possible directions and improvements. Thank you for making lab an accommodating and friendly environment conducive to the learning and exchange of ideas.

Chapter 1, in part, have been submitted for publication of the material as it may appear in Biomedical Circuits and Systems Conference, 2015, Sun, Alexander; Au,

Anthony; Venkatesh, A.G.; Gilja, Vikash; Hall, Drew A.,IEEE, 2015. The thesis author was a coauthor of this paper.

Chapter 2, in part, have been submitted for publication of the material as it may appear in Biomedical Circuits and Systems Conference, 2015, Sun, Alexander; Au, Anthony; Venkatesh, A.G.; Gilja, Vikash; Hall, Drew A.,IEEE, 2015. The thesis author was a coauthor of this paper.

ABSTRACT OF THE THESIS

Simulation of a Scalable Electrochemical Immunosignaturing Biosensor Array for Disease Diagnosis

by

Anthony Au

Master of Science in Bioengineering

University of California, San Diego, 2015

Professor Vikash Gilja, Chair
Professor Gert Cauwenberghs, Co-Chair

The ability to detect diseases during early progression greatly impacts the effectiveness of treatments, especially for cancers. Current research focuses on discovering specific biomarkers associated with the disease, but these are difficult to discover and present in very low concentrations. Conversely, immunosignaturing leverages the immense amplification provide by the immune system to examine antibody patterns on a random array of peptides. This thesis explores a new electrochemical detection method instead of the traditional optical detection for use with the immunosignaturing chip. I

developed a software simulation in order to investigate the parameters of the system. The first part of the simulation tracked the concentration changes of the molecules of interest. These concentrations of molecules drove the electrochemical discharges modeled in the next part of the simulation. The efficacy of the simulated discharges was determined by comparison with experimental discharge data. The first two parts of the simulation showed that crosstalk occurs with adjacent non-active sensors and the time delay before it happens largely depends on array geometry and sensor capacitance. The last part of the simulation explored the ability of this method to discern various diseases. Classification of a transformed optical data yielded similar classification accuracy compared to the original optical data. A mock end to end simulation demonstrated high accuracy as well. This thesis outlines a few approaches for implementation of the physical device, while laying out the framework to further explore parameter variations and disease classification.

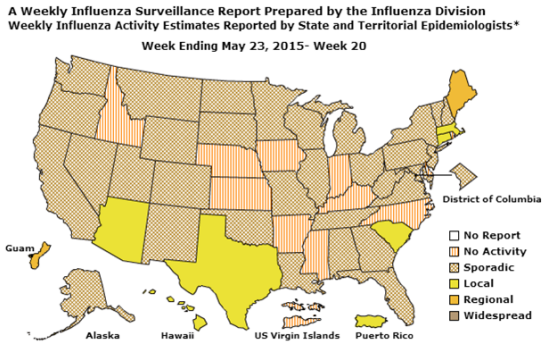
Chapter 1

Immunosignaturing

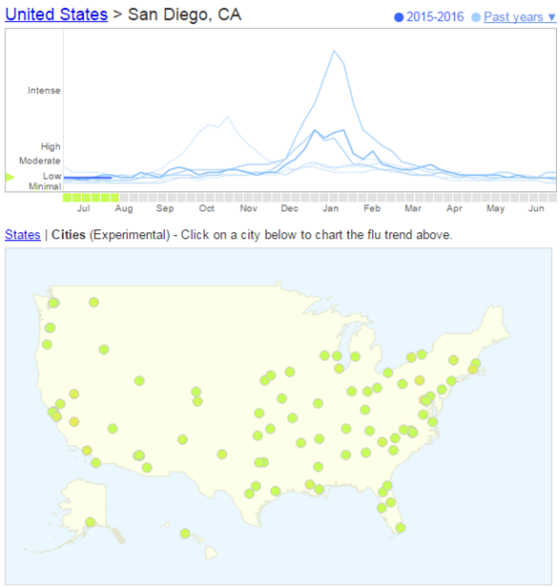
1.1 Introduction

Early detection of disease allows for earlier diagnosis and treatment, resulting in a higher chance of recovery than treatment at later stages [1]. This may be especially useful for diseases such as cancer, where the most effective treatments are performed in the earlier stages before metastasis occurs [1]. Early detection is not limited to improving patient treatment and recoveries but may also be used a method to track the spread of contagious infections, such as influenza (flu). Thus enabling health officials to establish preventative measures in the appropriate locations to stop or slow epidemics. Methods for tracking influenza have been employed by the Center for Disease Control (CDC) and Google Flu Trends. The CDC flu tracker relies on voluntary submissions of reports from United States health care related institutions, departments, and clinics [2]. Using these reports the CDC is able to determine what flu strands are circulating, changes in the flu virus, any flu related illnesses, impact of the flu on hospitalizations and deaths, and when and where flu activity is occurring[2]. The CDC does not predict where the flu will spread or appear, it merely monitors where it has occurred to allow authorities to take

appropriate measures. By this time though, the disease may have already spread further making it difficult and expensive to prevent as a wider net needs to be drawn. Google Flu Trends, which was discontinued on August 20, 2015, relied on tracking Google searches of flu or flu-related terms[3]. They may include things such as runny nose and sore throat, or they may be more obscure such as gloomy and sad. Google flu trends was designed to be independent from the CDC tracker and predict future outbreaks, but the reliability of search terms as an indicator is questionable due to human variability. These method do not use early detection and rely more on symptomatic features to track and predict disease outbreak and spread.



(a) CDC flu tracker



(b) Google Flu Trends

Figure 1.1: Current Flu Tracking Implementations

A rapidly growing area of research involved with early detection of diseases has been focused on finding specific biomarkers (RNA, DNA, proteins, peptides, or antibodies) correlated with a specific disease. Biomarkers have been heralded as the key to more personalized medicine and reduced healthcare costs [4]. For example, the implementation of a multiple inflammatory biomarker for detection of myocardial

infarction and ischemic stroke was estimated to save one million people a total of $187 million over five years, excluding testing costs. [5]. Rapid growth in this technology, fueled by advances in massively parallelized technologies such as DNA microarrays and proteomics, has led to the publication of over 150,000 scientific papers claiming the discovery of new biomarkers [6]. Even with this enormous research effort dedicated to biomarker discovery and voluminous papers published, fewer than 100 biomarkers have been approved for clinical use by the FDA [6]. Biomarkers also tend to be dilute and difficult to find. For instance, early cancer cells may only consist of several million cancer cells. If each cell released 1000 molecules of biomarkers into 5 liters of blood, it would result in a concentration of only $3x10^{-14}M$ [7]. As a result of the dilution in blood, it may take several decades to even reach detectable levels [8, 9, 10]. This exemplifies the difficulty of using biomarkers for disease detection.

Antibodies on the other hand can be present in the blood at high concentrations. This is attributed to the extraordinary amplification effect of the B-cell, which can produce 5000-20,000 antibodies a minute [11, 12] and replicate itself every 70 hours [13]. Considering a maximum lifespan of 4.5 months [14, 15], the B-cell can approximately produce a 10^{11} signal amplification in a single week. The B-cells produce antibodies when foreign objects in the body are detected by the immune system. The fact that antibodies are abundant and naturally produced in response to diseases, makes it an ideal molecule for detecting diseases, especially during the early stages. Even cancers will release biomolecules unfamiliar to the immune system [16, 17, 18], which in turn induces an immune response and antibody production [19, 20, 21]. But it is difficult to use antibodies as a biomarker because a given disease may have multiple epitopes that several or even hundreds of antibodies bind to [22]. A way to account for this non-specific binding, is the use of a large array of randomly generated peptides to capture many types of antibodies. This technique is called immunosignaturing, which allows the discovery of

disease patterns based on the various antibody binding affinities to the random peptides.

1.2 Current Immunosignaturing Technique

Immunosignaturing was developed by Phillip Stafford and Stephen Johnston at Arizona State University (ASU). This novel technique examines the body's immune reaction to diseases instead of detecting specific biomarkers associated with a particular disease. Since immunosignaturing focuses on antibodies, it is able to take advantage of the massive amplification by the immune system, which results in much higher concentrations present as opposed to the minuscule amount of biomarkers present [23]. The immunosignature does not rely on identifying specific antibodies as biomarkers for disease, but rather, it captures the whole profile of antibodies within the immune system. Therefore it is possible that several diseases are present at the same time, which makes it more difficult to distinguish the specific pattern associated with a given disease. To complicate matters, peptides that have low affinity to an antibody can still capture it by virtue of avidity caused by the tightly packed peptides [7]. As a result, diseases can be detected using classification algorithms, but little is known about the sequence information. Recently more effort has gone into deciphering the exact binding sequences and affinities [24].

1.2.1 Assay

The immunosignaturing assay consists of fixing peptides to a solid surface such as a glass or silicon backing. Peptides are advantageous over whole protein due to the ability to mass synthesize overlapping sequences for adequate coverage and stability since folding into tertiary structure is not necessary [22]. Peptides sequences were chosen to be 17 amino acids plus a 3 residue linker chain in order to allow for greater

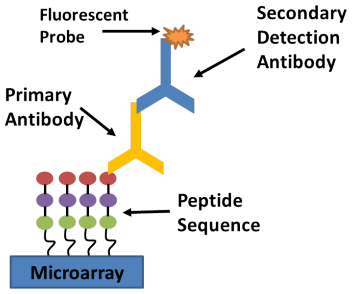


Figure 1.2: Illustration of antibody binding and detection to peptide array using optical detection

structural complexities [7]. Furthermore, the peptide sequences are randomly generated to give the device the ability to generalize for any type of disease [7, 23]. This can help to reduce costs as a single immunosignaturing chip has the potential to diagnose many diseases, making it unnecessary to manufacture a separate chip for each particular disease. Antibodies from a sample of blood serum then bind to the peptides creating a distinguishable pattern that can be classified with high accuracy [23, 25, 26]. The Fc domain (stem part) of the primary antibody is constant within the same class of animals, which allows a single type of secondary antibody to be used for detecting many types of primary antibodies. Figure 1.2 illustrates the binding of the primary antibody to the peptide sequence. A secondary antibody is then attached to the primary antibody for detection.

1.2.2 Optical Imaging

The current implementations of immunosignaturing analyze the binding patterns by using optical imaging to detect the relative antibody binding affinities. Peptides are bound to individual isolated wells with conformational and linear epitopes present in each well. In order to use optical imaging, a fluorescent probe is bound to the secondary detection antibody. After completing all necessary washes to eliminate weakly bound or free antibodies, the immunosignaturing array is inserted into an optical laser scanner.

The scan then sweeps a laser, set to the excitation wavelength of the fluorescent probe, over the sample and fluorescent intensity values are recorded. The scanner allows quick and easy reading of the immunosignature array, but can cost several thousand to tens of thousands of dollars and also tend to be quite bulky. Although, immunosignatures have a large potential cost savings over present diagnostic methods, these scanners are expensive; not every clinic may be able to afford them, especially poorer regions of the United States and the world, and it is certainly not affordable for the average household. This means serum samples have to be sent to either a central location or larger hospitals, which increases the time until a diagnosis is given to the patient.

1.3 Electrochemical Detection Method

Instead of using optical imaging, where an expensive and large scanner is necessary, electrochemical detection can be used for a small, rapid, and cost effective immunosignature device. The proposed electrochemical detection method aims to leverage a technique known as coulostatic discharge coupled with electrochemical amplification of interdigitated electrodes (IDE). These methods allow the fluorescent probe to be substituted for an enzyme capable of producing redox molecules (molecules that undergo reduction and oxidation in a reversible cycle).

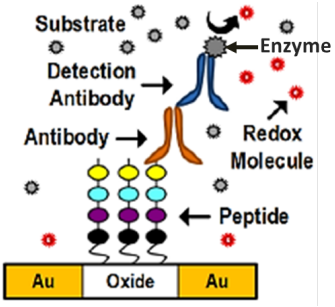


Figure 1.3: Illustration of antibody binding and detection to peptide immunosignature array using electrochemical detection

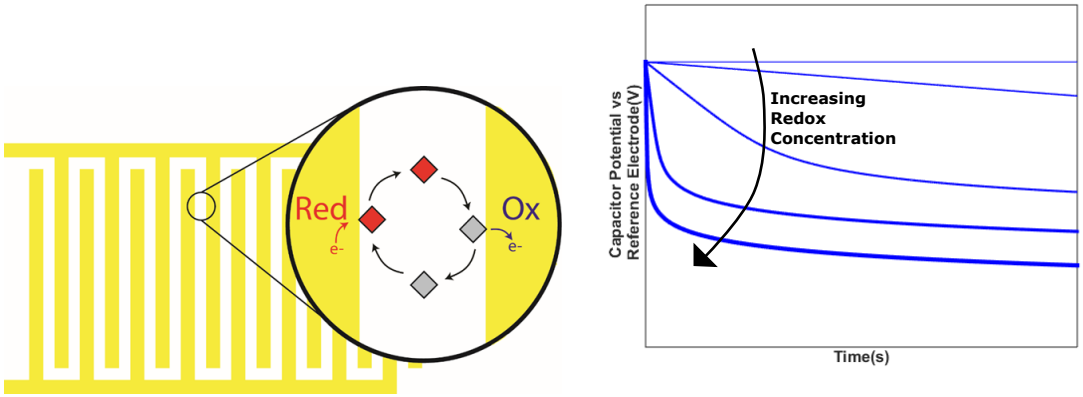
1.3.1 Coulostatic Discharge

Coulostatic discharge was discovered by Delahay and Reinmuth in 1962, where they described it as charging an electrode to a known quantity and observing the subsequent discharge of the double layer capacitance [27, 28]. The double layer capacitance is formed when the electrode in solution becomes charged at a particular potential. When the electrode is charged, either an excess or deficiency of electrons will be present in a thin layer (<0.1Å) on the surface of the metal electrode depending on the potential with respect to the solution [29]. The first layer, called the inner Helmholtz plane, where solvent molecules are specifically adsorbed and form a layer against the surface of the electrode; the second layer, called the diffuse layer, is composed of the non-adsorbed ions and extends into the bulk of solution [29]. In essence, the surface of the electrode and the oppositely charged ions in the solution form a capacitor with the adsorbed solvent acting as a dielectric. Coulostatic discharge leverages this technique to charge the double layer capacitance and measure its discharge over time once the charging potential is switched off.

1.3.2 Redox Cycling

While the resulting signal from coulostatic discharge was fairly large, additional amplification can be obtained through electrochemical means. First, two electrodes were used for a single sensing unit. Figure 1.4a shows that these working electrodes (WE1 and WE2) were configured as planar interdigitated electrodes. Second, a reversible redox pair was introduced into the system. The two working electrodes were then biased to the reduction and oxidation potentials of the specific redox molecule. Initially the biased potentials were held, but when the charging potential for one of the electrodes was released, the redox molecules cycle alternatively between the reduction and oxidation potentials as

the zoom in of figure 1.4a illustrates [30, 31]. The electrochemical amplification takes advantage of this ability to reversibly cycle between oxidized and reduced states, while shuttling electrons between the two electrode potentials. Figure 1.4b shows that higher concentrations of redox molecules lead to faster discharge rates, which was demonstrated by Zhu, Choi and Ahn [32].



(a) Redox cycling of an interdigitated electrode (b) Discharge rates as concentration increases

Figure 1.4: Electrochemical detection method

This amplification effect can be used to distinguish between different binding affinities because the enzyme substituted for the fluorescent probe is able to react with a substrate to produce redox molecules. Since the peptides are potted on top of the IDEs, the more antibodies that bind to that type of peptide on the IDE the more enzymes there will be bound nearby. This means that peptides with more bound antibodies produce more redox molecules, leading to faster discharge rates. This property allows electrochemical detection to be used for immunosignaturing. In addition, the electrochemical detection method captures time varying data, which may provide additional distinguishing features for disease diagnosis. The time varying data arises from the changing concentration at each sensor due to both diffusion and production of redox molecules. Instead of a single feature point such as the slope over a time interval, the time varying data contains many points. These points can be used to provide extra information as additional features, or

the sequential order of the sequence can be used to determine temporal patterns in the data.

1.3.3 Assay Sensitivity and Specificity

The sensitivity used here refers to the ability of the method to detect small amounts of antibodies, while specificity refers to the ability of the assay to distinguish between low and high antibody binding affinities. Compared to the current techniques enzyme-linked immunosorbent assay (ELISA) and phage display, both of which detect antibodies, immunosignaturing has greater specificity for low affinity interactions [33]. This specificity is due to the high density of peptides that trap peptides through avidity and rapid rebinding [7]. The binding of low affinity antibodies demonstrates the advantage of immunosignaturing over typical antibody detection methods, but the specificity should be the same between the optical detection and electrochemical detection, since they both rely on the same underlying antibody binding mechanism.

In terms of sensitivity though, the electrochemical detection method is greater than the optical detection method. The reason being that the optical detection method only has one fluorescent tag per antibody as seen in figure 1.2. On the other hand, in figure 1.3 one enzyme bound to an antibody can produce numerous redox molecules, resulting in an amplification of the original signal.

1.4 Proposed Chip

1.4.1 Bi-potentiostat

The bi-potentiostat circuit is used to record the IDE potential changes over time. The bi-potentiostat consists of three types of electrodes, the working electrode, the

reference electrode, and the counter electrode. The electrochemical redox reactions occur at the working electrode, causing changes in potential which can be observed and recorded. For a bi-potentiostat, there are two working electrodes, which for this system make up the components of the IDEs. The reference electrode is kept at a constant potential, which allows the changing working electrode to be measured relative to the controlled reference potential. The counter electrode allows the reference electrode to stay at constant potential by passing current through it instead of the reference electrode.

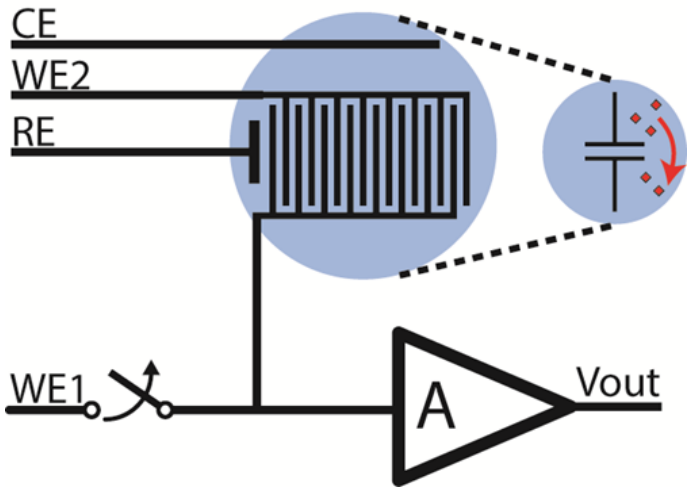


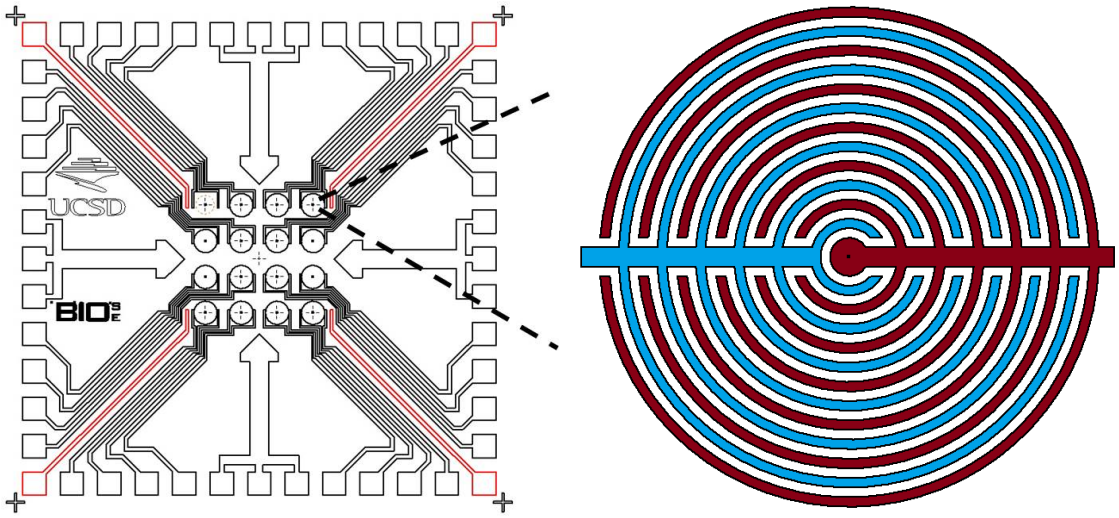
Figure 1.5: Three electrode bi-potentiostat

The two working electrodes in figure 1.5 serve as the two arms of the interdigitated electrode. The WEs are biased to the reduction and oxidation potentials to form the electric double layer for coulostatic discharge. After sufficient time has passed and the ions in the solution have reached a steady state at the electrodes, the switch connecting WE1 is opened to allow changes in potential. The substrate is then added to the solution and redox molecules are produced through reactions facilitated by the enzyme. The potential changes of WE1 are measured relative to the reference electrode which is typically set to ground. In the full array, the counter electrode and reference electrode will be shared by all of the sensors. Measurements of each sensor will be taken individually

by rapidly switching through the array via a multiplexer (mux).

1.4.2 Preliminary Sensor Array

A 4x4 IDE sensor array was initially fabricated to ensure that the electrochemical sensing performed as expected and produced distinct readings at each sensor. The whole chip measured 5mm in length and width. The IDE sensors were circular with a radius of $100\mu m$ and spacings between the two WEs of $5\mu m$. The diagram in figure 1.6b shows two WEs arranged in a circular interdigitated format. Each color corresponds to an individual WEs and the number of rings will vary based on the spacing and finger widths of the WEs. For instance, figure 1.6b is drawn for an IDE with a spacing and finger width of $4\mu m$, but the fabricated electrode with a spacing and finger width of $5\mu m$ will only have nine rings (the outer 4 rings are removed). The spacing between IDE sensors was



(a) 4x4 Electrode Array

(b) 16x electrode design

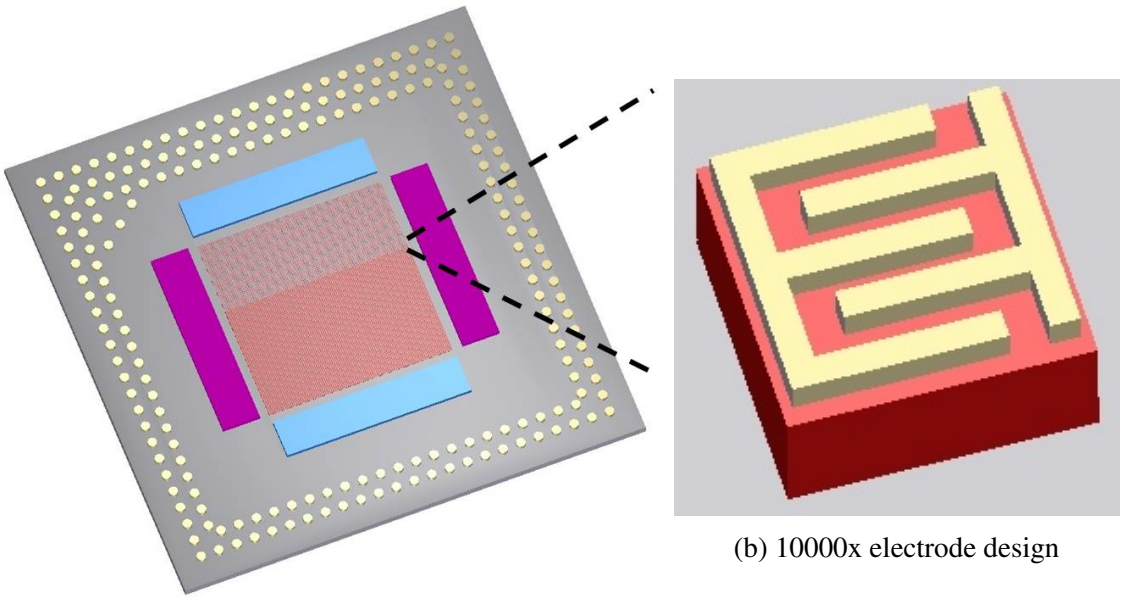
Figure 1.6: Preliminary 16 electrode array chip design

$283\mu m$ center to center. A 1.15mm radius o-ring was placed around the 4x4 array to form an impermeable boundary to contain the solution, which was filled to a volume of $20\mu L$. Peptides will be potted on the oxide layer between the two interdigitated electrodes with

a density of 1600-6400 peptides/μm^2. The peptides will protrude 20-30nm above the IDE sensors, which have a height of 100nm above the chip surface.

1.4.3 Scaled High Density Sensor Array

Once the discharge properties and readout circuits for the preliminary 16 IDE sensor array have been verified, the design can be scaled up to a higher density array. The proposed high density array will contain 10,000 sensors arranged in a 100x100 array. A rectangular electrode will be used due to ease of fabrication for the smaller dimensions and tightly packed grid. The sensor has proposed dimensions of $10\mu m$ length and width, a spacing and width of $11\mu m$, and center to center spacing between IDE sensors is $20\mu m$. This chip will also have a width and length of 5mm and an o-ring radius of 1.15mm.



(a) Conceptual 100x100 Electrode Array

(b) 10000x electrode design

Figure 1.7: High Density Electrode Array

These array dimensions may be changed based upon simulation results and experiments with the 4x4 array.

1.5 Challenges

Although electrochemical detection for immunosignaturing can lead to reduced costs and miniaturization, there are several challenges associated with both electrochemical detection and immunosignaturing that must be addressed for a viable device. Unlike the optical technique, which isolated different peptides within individual well plates, the electrochemical technique exposed each sensor to the same bulk solution. Since the electrochemical detection method relies on redox molecules freely floating in the solution, it is possible for them to diffuse across the device. This means that the discharge for a particular sensor may not be due to the antibodies binding to that sensor's peptides but instead, due to the diffusion of redox molecules from a neighboring sensor. This crosstalk between sensors adds an additional source of noise and may interfere with accurate classification of the discharge readings. The first part of this thesis develops

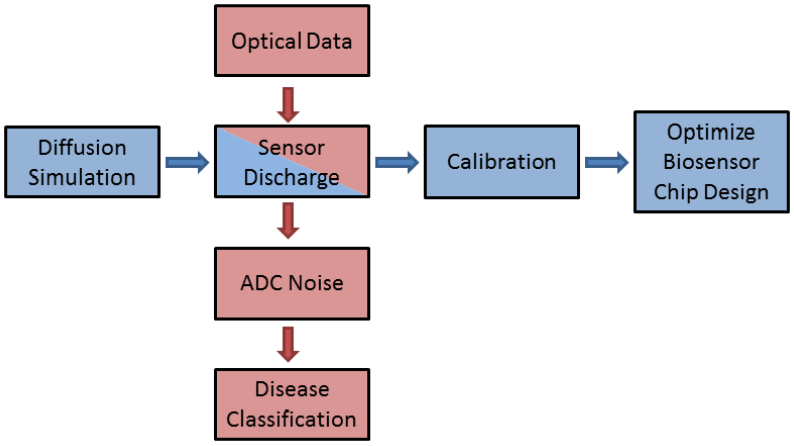


Figure 1.8: project pipeline

a simulation approach to examine this potential crosstalk. The blue path of figure 1.8 illustrates the separate steps taken to develop and use the simulation of the device physics to understand and optimize the parameters. This path will be addressed in chapter 2.

The red path of figure 1.8 illustrates the separate steps taken to simulate the end

to end performance of the device by starting with the optical data and transforming it into approximate sensor discharge values for disease classification. The major challenges for this path consist of transforming the fluorescent intensity values into corresponding discharge rates and the curse of dimensionality due to the high dimensional data and limited number of samples. These issues will be addressed in chapter 3 and chapter 4.

Chapter 1, in part, have been submitted for publication of the material as it may appear in Biomedical Circuits and Systems Conference, 2015, Sun, Alexander; Au, Anthony; Venkatesh, A.G.; Gilja, Vikash; Hall, Drew A.,IEEE, 2015. The thesis author was a coauthor of this paper.

Chapter 2

Simulation

2.1 Diffusion Model

Many methods have been developed to simulate the diffusion of particles. With these different methods comes a trade-off between modeling resolution and computational complexity. More complex methods such as the partial differential equation (PDE) for diffusion have better resolution than the compartment model. The compartment model consists of many discrete compartments, while the PDEs consist of continuous points over the entire are defined by boundary conditions. Both models are equivalent if the given comparment size equals the spatial grid size of the PDE. The compartmental model was chosen for speed and simplicity because the compartments only need to be as large as the sensor in order to obtain distinguishable discharge readings. In addition, it is easier to implement boundary conditions and physical alterations to specific compartments.

2.1.1 Compartment Model

The compartment model aims to divide the system into many individual compartments. Compartment models have applications in numerous fields including phar-

macokinetics and epidemiology. This model is considered a lumped element model because it lumps the spatial concentration distribution of each compartment into a single value, effectively simplifying the partial differential equation into a system of ordinary differential equations.

The compartmental model looks at how much charge is within one capacitor instead of the exact spatial distribution of charge within that capacitor. For the immunosignaturing simulation, the concentration of redox molecule within the entire compartment was examined instead of the distribution of particles within that compartment. Figure 2.1a depicts one compartment and its interaction with neighboring compartments. The compartment shown in figure 2.1a was a box; therefore it comprised of 6 separate faces for the flux of redox molecules. The amount of molecules traveling to different compartments is proportional to the current number of molecules in the originating compartment. For example, looking at the flow between compartments 1 and 7, a large number of molecules in compartment 1 will cause more collisions between molecules and result in the dispersion of more molecules towards compartment 7. Collisions are less likely for a small number of molecules in compartment 1, resulting in the dispersion of fewer molecules. The proportionality constant that controls this relationship is rate constant d_{17}.

Compartment Model Assumptions

Several assumptions were made about the real system in order to use the compartment model to approximate the diffusion of redox molecules from the reaction point. The following assumptions were either inherent to the compartment model or were made to simplify certain geometries.

1. The concentration within each compartment was uniform. This assumption simplified the partial differential diffusion equation by eliminating the spacial component, which resulted in an ordinary differential equation with respect to time. Instead, the

spacial component of the model was determined by the individual compartments, the size of which may be adjusted based on the tradeoff between finer resolution and computational efficiency.

2. The system was assumed to be open when redox molecules are actively produced and closed when there was only an initial impulse of redox molecules. The volume of each compartment was assumed to be constant over time. This was an important assumption for the concentration calculation as fluctuations in volume will vary concentrations. Volume changes can occur due to external factors that result in evaporation or leakage of serum. In either case, the most relevant bottom compartment layer closest to the sensor is unlikely to experience any changes.

3. The rate of material flow between compartments is directly related to the amount of materials within each compartment.

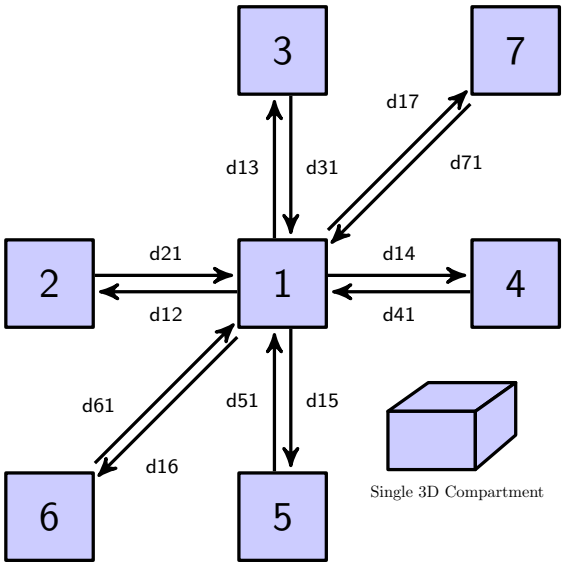
4. The rate constants were assumed to be equal in all directions. There were no external forces acting on the system so the redox molecules should only be under the effects of diffusion.

5. Each compartment has a box shaped geometry. The box shape compartment had the most organized 3-D structure. The disadvantage of a box geometry was the need for the particles to travel from an adjacent face instead of along the corners. This meant more time was needed for diffusion to compartments along the diagonals, which do not share a face with the original compartment. However, this issue can be resolved with finer resolution as smaller compartments captured the spacial changes more distinctly and better approximated the unrestricted movement of molecules.

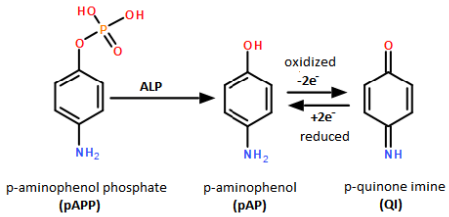
6. The o-ring, bottom surface, and air-serum interface were considered as no flux

boundaries. This meant no particle was able to travel past this boundary. Therefore the rate constants were zero at the border, which resulted in the rate of change in concentration for compartments outside of the border to be equal to zero.

7. The initial concentration of all compartments was equal to zero. At time = 0 seconds, there were no particles of the redox molecules within the system. They only appear after time = 0 seconds when the enzyme reacts with the substrate molecules introduced into the system.



(a) Single element of compartment model



(b) Redox cycle between redox molecules pAP and QI initiated by the reaction between substrate pAPP and enzyme ALP

$$\frac{dC_1}{dt} = Rxn(t) - d_{12}C_1 - d_{13}C_1 - d_{14}C_1 - d_{15}C_1 - d_{16}C_1 - d_{17}C_1 + d_{21}C_2 + d_{31}C_3 + d_{41}C_4 + d_{51}C_5 + d_{61}C_6 + d_{71}C_7 \tag{2.1}$$

The dimensions of the immunosignaturing chip were used to approximate a box with the length and width of the chip and the height of the fluid placed within the o-ring. This box encompassing the system was further subdivided into smaller boxes which constituted

the individual compartments. Since the compartments consist of boxes, there were 6 possible surfaces that the material can flux across. Figure 2.1a shows the interaction of compartment 1 with the 6 neighboring compartments.The arrows show the direction of flow and the labels represent the rate constants across compartments. These rates have units of $time^{-1}$ and were calculated by multiplying the diffusion constant of the relevant redox molecule with the surface area of each face on the compartment. Of course the units for the surface area must be the same as the diffusion constant or converted before calculating the rate constants.

The differential equation for system depicted in figure 2.1 can be constructed by adding or subtracting the flows to the particular compartment. Equation 2.1 depicts the differential equation for compartment 1 of figure 2.1a. The $Rxn(t)$ term represents the reaction between the enzyme and substrate that produces the redox molecule. For this device, Alkaline Phosphotase (ALP) and p-aminophenol phosphate (pAPP) will be used respectively for the enzyme and substrate. The reaction will produce the redox couple, p-aminophenol the pAP redox molecule. The reaction is shown in figure 2.1b. This will be discussed in the following section. All of the arrows pointing away from compartment 1 depict molecules flowing out of compartment 1 and are represent mathematically by subtracting the current concentration of compartment 1 multiplied by the rate constant for that particular arrow. Conversely, arrows pointing towards compartment 1 depict molecules flowing in and can be represented by adding the rate constant for that arrow multiplied by the current concentration of the compartment flowing towards 1.

2.1.2 Reaction Rate

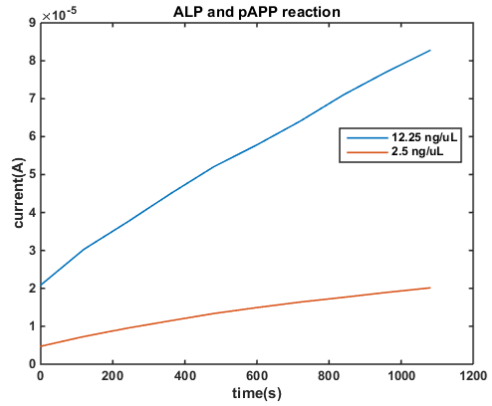
The $Rxn(t)$ term in equation 2.1 represents the rate at which the pAP was produced from the ALP-pAPP reaction. This parameter was estimated empirically from a voltammetry experiment, where an excess amount of pAPP was introduced into a system

with a fixed amount of ALP. Voltammetry consists of changing the working electrode potential and examining the produced current between the working electrode and the counter electrode in a 3-electrode potentiostat setup. The changing potential can cause redox molecules to undergo reduction or oxidation, which affects the current as charges are carried by these molecules. For this experiment, mass concentrations of 2.5 and 12.5 $\frac{ng}{\mu L}$ of ALP were used. The current at the working electrode was measured every 2 minutes with time = 0 seconds representing the time when pAPP was introduced into the system. Figure 2.2a shows the measured current for both concentrations of ALP over a period of 18 minutes. The slope for the $12.25\frac{ng}{\mu L}$ ALP was steeper, indicating the presence of more redox molecules.

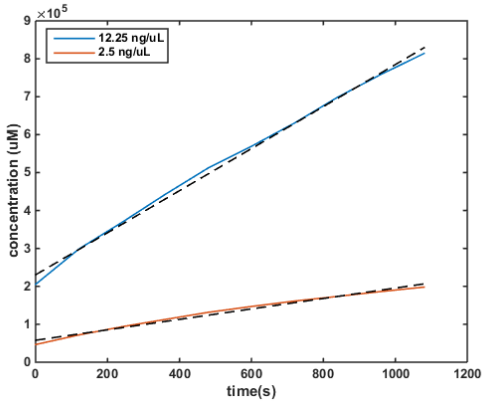
$$i_p = 0.4463nFAC(\frac{nFvD}{RT})^{\frac{1}{2}} \tag{2.2}$$

where n is the number of redox electrons transferred, F is Faraday's constant, v is the voltage scan rate, D is the diffusion coefficient for the redox molecule, R is the gas constant, T is the temperature, C is concentration, and A is electrode surface area. The Randles-Sevcik equation(eq 2.2) was used to relate the current from the voltammetry experiment to the concentration of pAP produced. Application of equation 2.2 leads to concentrations shown in figure 2.2b. Least squares linear regression was applied to the concentration over time curve for each amount of ALP. The reaction rate for each amount of ALP enzyme was determined by the slope of the regressed lines. The final transformation consisted of interpolating the reaction rates for any concentration of ALP, not only for the tested 2.5 and 12.25 $\frac{ng}{\mu L}$. A piecewise linear relationship was assumed between the at different ALP concentrations and zero reaction rate for zero ALP presence. The mass concentration of ALP was then converted into a concentration by dividing by the molecular weight with appropriate units and assuming an area of $1\mu m^3$, resulting in figure 2.2c. For the following simulations a single antibody to single peptide bond was

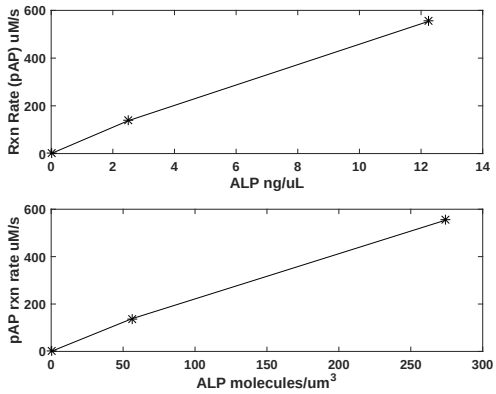
assumed, which translated to the concentration of ALP at a single sensor to be equivalent to the peptide potting density.



(a) current due to reaction



(b) Concentration transformation



(c) RxnRate per Density of ALP

Figure 2.2: ALP-pAPP reaction rate

2.2 IDE Sensor Discharge Model

After simulating the diffusion of redox particles, the next step was to calculate the resulting sensor discharge readings. Since the sensors consist of IDEs which function as a capacitor, the current-voltage relation for an ideal capacitor can be used to predict the discharge curves. Several assumptions were made in order to derive the relationship

between the potential and current. An extra calibration constant was used to adjust the curves to the bias of the second working electrode.

IDE Model Assumptions

1. *IDEs form an ideal capacitor*: This allows the basic capacitor equation relating current to the rate of change of potential to be used to simulate the discharge over time.

2. *Electrodes follow Nernstian properties*: Allows use of Nernst equation to determine concentration ratio of oxidized and unoxidized species.

3. *Linear concentration profile for Nernstian electrodes*: Allows the concentration differential to be simply equal to the difference between the reduced species concentration at each electrode assuming species originally exists in reduced form.

4. *Quasi-steady state electrochemical reaction*: This allows the time differential of concentration to be ignored. Instead many static concentration points will be simulated and pieced together.

2.2.1 Discharge Equations

The discharge of each IDE sensor can be described by the capacitor discharge equation.

$$-C_{extra}\frac{dE(t)}{dt} = I(t) \tag{2.3}$$

where C_{extra} is the capacitance of the IDE sensors, $E(t)$ is the potential of the capacitor/IDE, and $I(t)$ represents the electrochemical reaction current between the redox biased working electrodes.

Following assumptions 2 and 3, the electrochemical reaction current can be

described by:

$$I(t) = nFAD_R(\frac{C_{R,1}^{lim} - C_{R,2}(t)}{\delta_R}) = nFAD_O(\frac{C_{O,2}(t)}{\delta_O}) \tag{2.4a}$$

$$\delta_R = \delta_O = kW \tag{2.4b}$$

where n is the number of electrons during each redox reaction, F is Faraday's constant, A is the top surface area of an IDE sensor, D_R is the diffusion coefficient for the reduced species, D_O is the diffusion coefficient for the oxidized species,$C_{R,1}^{lim}$ is the saturated concentration value of reduced species at WE1, $C_{R,2}(t)$ is the concentration value of reduced species at WE2, $C_{O,2}(t)$ is the concentration value of the oxidized species at WE2, and δ is the thickness of the diffusion boundary layer. The thickness of the boundary layer is proportional to the separation distance between the two working electrodes, and the proportionality constant, k, depends on the geometry of the electrodes as shown in equation 2.4b above. The simulations performed in this thesis assume a δ value simply equal to separation between the two working electrodes.

The limiting current is the maximum current that results from the electrochemical reaction. All parameters in equation 2.4 are constants pertaining to the system except $C_{R,1}^{lim}$ and $C_{R,2}(t)$, which varies as the IDE is discharged. The maximum difference $C_{R,1}^{lim}$ - $C_{R,2}(t)$ occurs at the steady state. The working electrodes are biased to the redox potentials and it is assumed that all of the reduced species aggregate at the electrode biased to the reduction potential and vice versa for the oxidized species. Therefore, the concentration of the reduced species at WE1 will be maximum before discharge and the concentration of the reduced species at WE2 will be equal to 0.

$$I^{lim} = nFAD_R(\frac{C_{R,1}^{lim}}{\delta_R}) = nFAD_R(\frac{C_{R,1}^{lim}}{\delta_R}) \tag{2.5}$$

The Nernst equation predicts the ratio of oxidized to reduced species concentration at WE2 because not every single reduced species that diffuses from WE1 to WE2 when the potential drops is oxidized at WE2.

$$E(t) = E^{0'} + \frac{RT}{nF} ln[\frac{C_{O,2}(t)}{C_{R,2}(t)}] \tag{2.6}$$

where $E^{0'}$ is the standard reversible potential of the redox species. There is also a material balance at WE2, so the concentration of the reduced species plus the concentration of the oxidized species at WE2 equals the saturation concentration of oxidized species, $C_{O,2}^{lim}$.

$$C_{R,2}(t) + C_{O,2}(t) = C_{O,2}^{lim} \tag{2.7}$$

Solving equations 2.4 - 2.7 derives a relationship between potential and current.

$$E(t) = E^{0'} + \frac{RT}{nF} ln[\frac{I(t)}{I^{lim} - I(t)}] \tag{2.8}$$

Now equation 2.8 can be substituted into equation 2.3 in order to obtain an ordinary differential equation for the sensor potential.

$$1 + exp[\frac{nF}{RT}(E(t) - E^{0'})]C_{extra}\frac{dE(t)}{dt} + I^{(}lim)exp[\frac{nF}{RT}(E(t) - E^{0'})] = 0 \tag{2.9}$$

Equation 2.9 is derived from [30, 32]. However, it did not account for the bias due to the oxidizing potential applied to WE2. The sensor potential is the difference in the potentials of WE1 and WE2 which means that it will stop discharging when that difference becomes 0. This happens when the potential of WE1 falls to the potential of WE2, and since the sensor potential is measured with respect to the reference potential, the measured potential will equal the WE2 potential minus the reference potential. In

order to correct for the bias, equation 2.9 is manipulated to solve for the potential $E(t)$.

$$E(t) = \frac{RT}{nF} ln(\frac{-C\frac{dE}{dt}}{C\frac{dE}{dt} + I^{lim}}) + E^{0'} \tag{2.10}$$

The limit of equation 2.10 when the discharge rate $\frac{dE}{dt}$ equals 0 is taken to determine the potential where discharge stops. The limit approaches $-\infty$ meaning the discharge of the sensor never ceases. A constant α is added inside the log term of equation 2.10 in order to prevent a log(0) term resulting in the $-\infty$ limit. The potential is set to WE2 bias potential, E^{W2} because the limit when $\frac{dE}{dt}$ goes to 0 should equal the WE2 bias potential as discussed above. Since $\frac{dE}{dt}$ equals 0 in the limit, it is possible to solve for the calibration constant α using simple algebraic manipulations.

$$E(t) = \frac{RT}{nF} ln(\frac{-C\frac{dE}{dt}}{C\frac{dE}{dt} + I^{lim}} + \alpha) + E^{0'} \tag{2.11a}$$

$$\alpha = e^{\frac{nF}{RT}(E^{W2} - E^{0'})} \tag{2.11b}$$

Finally, the corrected sensor discharge equation is discerned by substituting equation 2.11b into equation 2.11a and applying algebraic manipulation to obtain the same form as the uncorrected discharge equation(eq. 2.9). The simulated sensor capcitor potential now discharges to the WE2 bias potential given sufficient time.

$$[1 + e^{\frac{nF}{RT}(E(t) - E^{0'})} - e^{\frac{nF}{RT}(E^{W2} - E^{0'})}] C_{extra} \frac{dE}{dt} + I^{lim} [e^{\frac{nF}{RT}(E(t) - E^{0'})} - e^{\frac{nF}{RT}(E^{W2} - E^{0'})}] = 0 \tag{2.12}$$

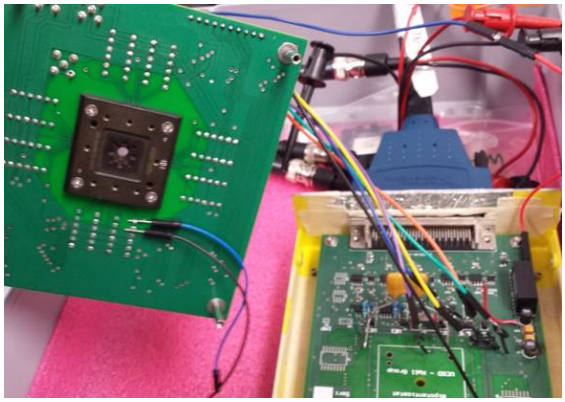
2.3 Preliminary Experiments

The previous section demonstrated a method to simulate the discharge curve of an IDE sensor, but there was no evidence that the physical device followed the predicted

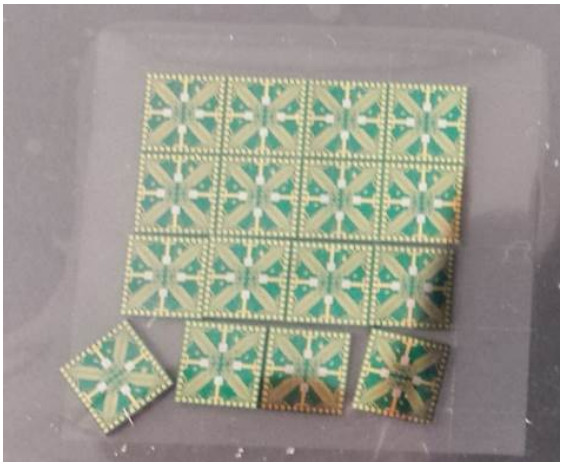
trajectory. An experiment was conducted with a 16 sensor array in order to ascertain the validity of the simulated trajectories. The experiment was performed by Alexander Sun of Professor Drew Hall's lab.

2.3.1 Test Setup and Parameters

The initial setup was not properly equipped for a full reaction+diffusion to sensor discharge. Consequentially, the resulting measurements were only comparable to the sensor discharge simulation. One of the chips shown in figure 2.3b was inserted into the board on the left of figure 2.3a with a small o-ring placed in between to surround the sensors. PBS was used as the buffer solution to substitute for blood serum. In addition, ferri/ferrocyanide was used as the redox molecule due to the convenience of availability. The diffusion coefficient used for ferricyanide was $7.0x10^6 cm^2/s$. The redox reaction between ferri/ferrocyanide only transferred 1 electron, so "n" in equation 2.12 had a value of 1 instead of 2.



(a) Preliminary device setup with slot for chip and multiplexer array



(b) Fabricated test sensors with spacings between working electrodes of 5um. Each chip contains a 4x4 array of IDE sensors as depicted in figure 1.6b

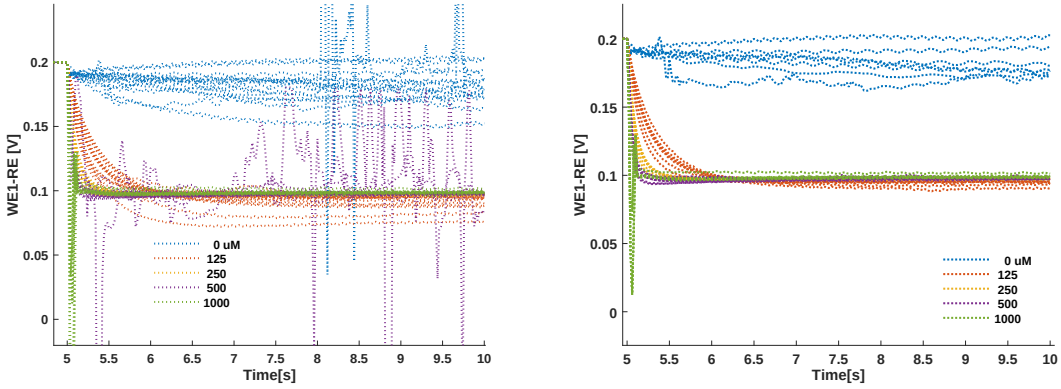
Figure 2.3: Preliminary test setup

In the ideal system peptides will be potted onto the surface of the sensors, allowing bound enzymes to react and create the redox molecules that facilitate the sensor discharges. The system will also be able to monitor the discharge rates of all 16 sensors during the same testing period (i.e. measurements will be made by rapidly switching between sensors without the need to reset for each sensor). However, due to time and resource constraints, the redox molecules were directly pipetted into the solution surrounding the chip. The concentrations of pipetted redox molecules varied from 0, 125, 250, 500, to 1000μM. Furthermore, the mux hardware was not fully functional, making it necessary to perform individual tests for each sensor. Not only was this more time consuming, but the diffusion of redox particles across the chip could not be measured and compared against the diffusion simulation. The sensor discharge model was the most uncertain modeling diffusion concentrations, which has been well established in other fields including pharmacokinetics [34]. The dimensions of the chips were the same as the 16 electrode array described in chapter 1, and all measurements of the working electrode 1 potential were made against a reference electrode set to 0V. Initially, the working electrodes were charged to a potential of 0.2 and 0.1V (reduction and oxidation potentials for ferri/ferrocyanide). The charging source of 0.2V for WE1 was switched off at five seconds, and discharge of the sensors occurred afterwards.

2.3.2 Calibration and Results

The experimental results followed the expected trend of higher discharge rates when higher concentrations of pAP were used, as shown in figure 2.4a. The individual lines within one color represent measurements for each of the 16 different sensors on a single chip. The measurements were very noisy and there was a current leak, which contributed to the large dips in the curves for 500 and 1000 μM redox concentrations. It can also be evidenced from the IDE discharges when no redox was added to the system

($0\mu M$). Ideally for 0 μM, no redox molecules should be present to create the current between the two working electrodes, which would lead to no discharge, but the blue curves in figure 2.4a show discharges in the IDE sensor from the initial charged potential of 0.2V. The initial discharge rates were calculated by finding the slope of each curve right after discharge began at five seconds. Figure 2.4b shows the initial discharge rates of non-defective or "good" sensors only. Sensors were determined to be defective if the discharge curves demonstrated erratic behavior. The discharge curves depicting only the good sensors were much smoother than the discharge curves of all sensors including the defective ones.



(a) Measured discharge curves for all 16 sensors. Each color represents a test using a different substrate concentration.

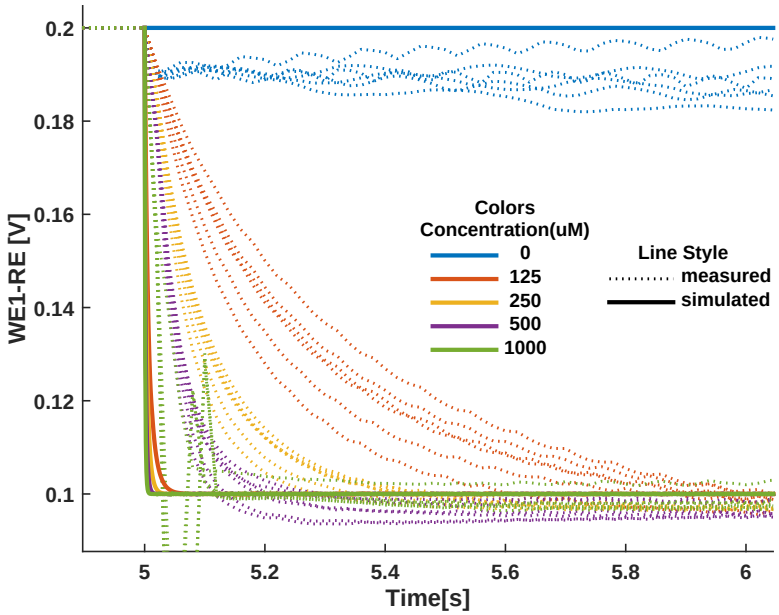
(b) Measured discharge curves for 7 non-defective sensors.

Figure 2.4: Experimental Results

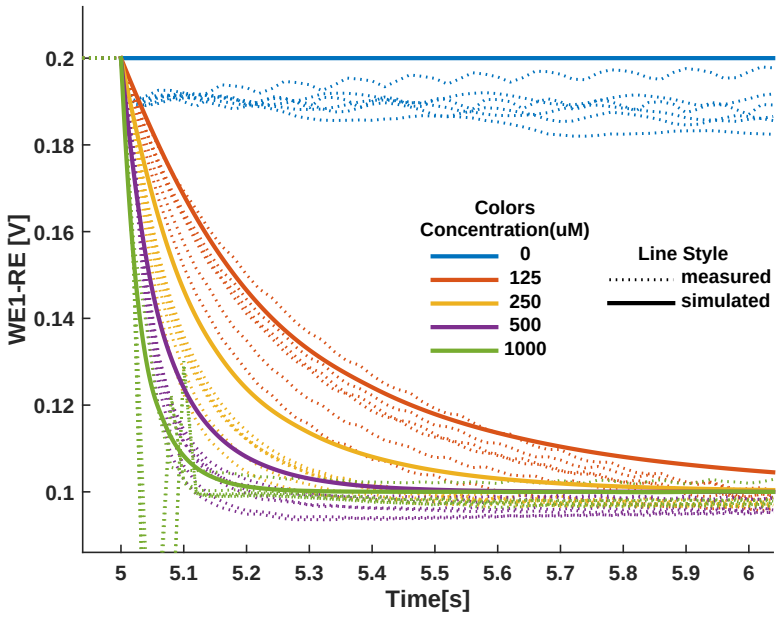
The simulated discharge curves were compared to the measured curves using the physical parameters of the chip and sensors as well as the redox molecules. The diffusion boundary thickness was assumed to be equal to the separation between working electrodes and the experiment was performed at room temperature(296K). The impedance of the electrochemical cell (electrode plus PBS solution) was measured with electroimpedance spectroscopy and fit to a Randles circuit model, which contains a double layer capacitor in parallel with the impedance of a faradaic reaction. The capacitance calculated through

this method was $2.2699x10^3 pF$.

Figure 2.5a shows the simulated curves based on the parameters above and the capacitance of $2.2699x10^3 pF$. The discharge curves are much steeper than the measured curves and completely discharge within 50ms. This difference between the simulated and measured results is either due to limitations of the model or incorrect values for parameters. Since the simulated curves still show the same discharge shapes and follow the trend of higher redox concentrations leading to higher discharge rates, it is simpler to add a calibration factor to the discharge model. This is fine for the purpose of the simulation, since the goal is to predict the discharge values of the real device not necessarily to understand the complete physical model behind the device. Although this means that at least one IDE sensor with appropriate dimensions will need to be fabricated for calibration purposes, it prevents the need fabricate all 10,000 or more sensors before optimizing chip dimensions and parameters. For discharge equation 2.12, the most uncertain parameters were the diffusion boundary thickness and the capacitance of the electrode sensor. Changing the diffusion boundary layer affects the I^{lim} term, so a smaller diffusion layer will decrease the discharge rate and vice versa. However, the diffusion boundary thickness was already assumed to be equal to the distance separating the two working electrodes when in reality it is likely much smaller and closer to the electrode surface. The capacitance measurement, on the other hand, was more likely to vary due to assumptions made when fitting the impedance data to a circuit model. The capacitance of the electrode sensor was inversely proportional to the discharge rate, a higher capacitance led to a smaller discharge rate. A calibration scaling factor for the capacitance was determined by sweeping through several values and selecting the scaling factor with the lowest error in reference to the mean measured discharge rates. The error was determined by simply summing the absolute value of the difference between the simulated initial discharge rates and the mean measured rates. The final calibration



(a) Comparison of uncalibrated simulation curve and measured curves



(b) Comparison of calibrated simulation and measured curves

Figure 2.5: Measured Curves versus Simulated Curves

factor was 42.2 times the original capacitance of $2.2699x10^3 pF$, and it had an error of 0.8407 V/s compared to 5.92 V/s in the uncalibrated simulation. The simulated curve

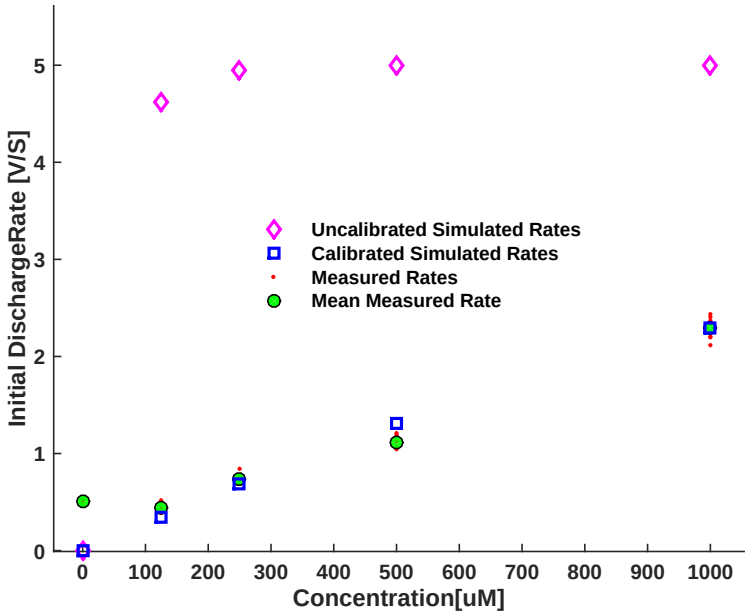


Figure 2.6: Calibrated and uncalibrated initial discharge rates compared to measured rates

after calibration in figure 2.5b closely followed the measured curves.

2.3.3 Discussion

The preliminary experiment showed that the sensor discharge simulation can closely follow the actual measured discharge curves. The longer time measurements were noisier, but most of the information will be extracted within a few seconds to a minute due to the effect of crosstalk. The calibration was factor of 42.2, but the calibrated capacitance was actually closer to the suspected value for the size of the sensor. The deviation between the measured capacitance and the calibrated capacitance may be attributed to errors during the capacitance measurement process or an assumption of the model that poorly predicts the real physical system. In the future, a finite element analysis can be performed to simulate the IDE sensor within electrolyte to determine a more accurate value of capacitance. Nevertheless, the most important result was that the

simulated curves were able to predict the measured curves with one calibration parameter. This will prove important when simulating the high density sensors and will give a high level of confidence in the prediction of the simulated results. In the future, the next step will be to add the peptides and monitor the diffusion of redox molecules throughout the system.

2.4 Crosstalk Analysis

As discussed in chapter 1, crosstalk between the nearby sensors can introduce noise into the discharge measurements, which in turn can affect the disease classification performance. Simulations were performed on both the preliminary and high density array. For simplification, it was assumed that redox molecules were produced at the rate described in section 2.1.2. A central location was observed to examine the effects from crosstalk on the discharge rate of the central location. Each location in figure 2.7 represented an IDE sensor. For instance, sensor 6 in figure 2.7 was chosen as the central location for the 4x4 preliminary sensor array. The initial test only had peptides placed at sensor 6, but in subsequent tests, peptides were placed in sensor 11 and 16. Only sensor 6 had peptides attached in every test; other sensors were only present for a single test. For

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

Figure 2.7: Peptide attachment locations for crosstalk analysis

example, the first test had peptides placed only at sensor 6, the second test had peptides placed at sensors 6 and 11, and the final test had peptides placed at sensors 6 and 16. This was done to examine the effect of distance on crosstalk noise at sensor 6. It was expected that the crosstalk effect from sensor 11 would be larger than the effect from sensor 16 due to the longer distance over which diffusion must occur.

2.4.1 Preliminary Sensor Array Analysis

For the crosstalk analysis of the 4x4 preliminary sensor array, the peptide locations were assigned as discussed in section 2.4. The simulation was run using the reaction rate for ALP and pAP derived in section 2.1.2 because empirical measurements were only performed for the ALP and pAPP enzymatic reaction. The sensor capacitance value used for this analysis was the calibrated value from section 2.3.2, which was obtained by using ferricyanide instead of pAP as the redox molecule. Although the redox molecules switched from ferricyanide to pAP, the sensor capacitance should not change because the double layer depends on the ions within the buffer solution and any transverse capacitance result from the chip and electrode materials as well as dimensions. The double layer capacitance is a function of potential [29], but the capacitance measured did not vary significantly within the 0-0.3V range, which is close to the reduction potential of 0.35V for pAP. The other device dimensions were the same as the dimensions used for the calibration in section 2.3.2.

If the diffusion of redox molecules from sensors 11 and 16 affected the discharge curves at sensor 6, then one would expect the red and yellow curves for sensor 6 to have a faster discharge rate (slope). The discharge curves for these three tests behaved contrary to this expectation, which indicated that crosstalk did not affect sensor 6. The time needed for diffusion to the next diagonal is shown by the blue curve for sensor 11 in figure 2.8. Although the sensor with actively produced redox molecules was not affected

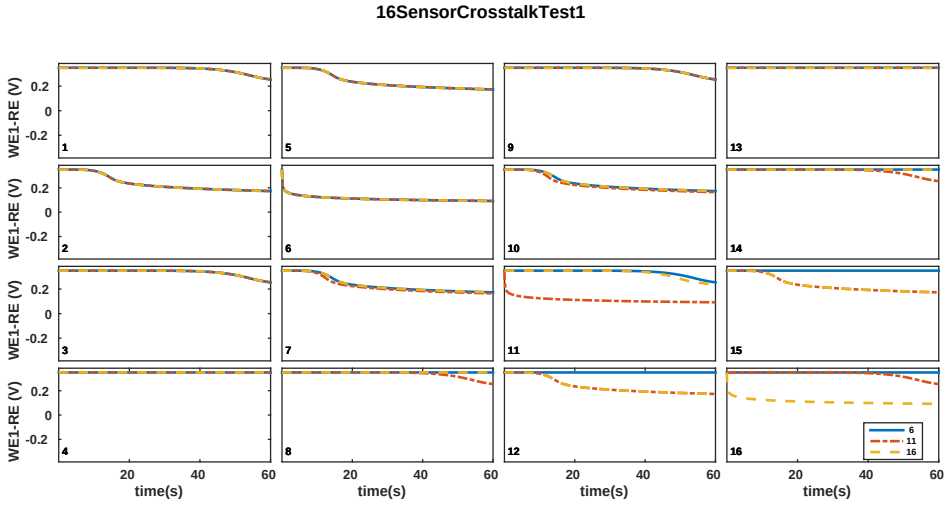


Figure 2.8: Crosstalk for 16 sensor array

by neighboring crosstalk, the nearby non-active sensors (e.g. sensor 7) showed a change in the working electrode potential. Since redox molecules were not produced at these sensors, the expected concentration was $0\mu M$. Based on the simulated curve for $0\mu M$ as seen in figure 2.5, the potential was not expected to change. Yet since there was a change, it must be due to crosstalk from the nearby active sensors. For the preliminary 16 sensor array, crosstalk required approximately 10 seconds to occur, which gives a rather large time frame to extract trustworthy measurements.

2.4.2 High Density Sensor Array Subsection Analysis

After examining crosstalk for the smaller 16 sensor array, crosstalk for the larger proposed 10,000 sensor array is simulated. In order to simulate the full 10,000 sensor array, a very fine resolution was needed to have each sensor modeled by at least one compartment. If the compartments encompass more than one sensor, it becomes impossible to distinguish concentration differences between the sensors due to the assumption of uniform concentration within a single compartment. A simulation for all 10,000 sensors though, would likely take several months to complete due to the sheer number of

compartments. The computational load can be reduced in several ways: 1) by increasing the time step of the ordinary differential equation solver used for the diffusion and sensor discharge equations, 2) by reducing the simulated solution height, thereby reducing the volume of the system, or 3) by simulating a small subset of the array, since it will take too long for enough redox molecules to cause crosstalk by traveling from one corner of the array to the opposite corner. The first option can result in the loss of important information if crosstalk occurs within a few ms or tens of ms. The second option will effectively reduce computational load, since less elements can achieve the same resolution and the most important area is the sensors on the bottom layer. However, this approach will still require a significant load since the discharges of 10,000 sensors need to be computed. The final option is the most computationally efficient as it simulates both a smaller area and number of sensors. Initial crosstalk occurs at neighboring sensors with limited effect on distant sensors, which makes simulating a subset of the 10,000 sensors to examine local diffusion and crosstalk a viable approach.

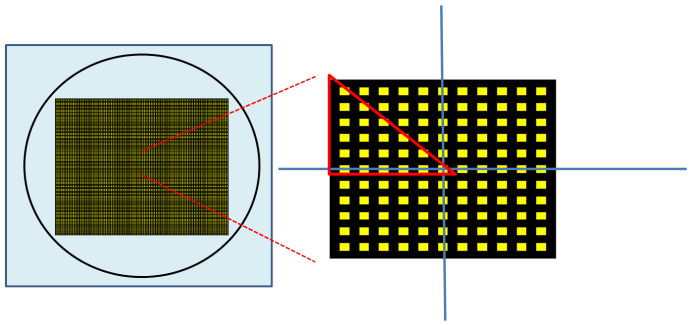


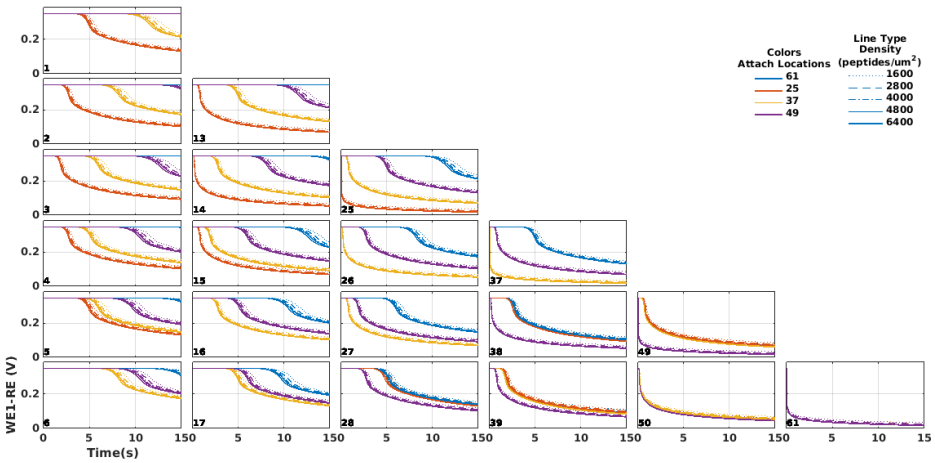
Figure 2.9: 11x11 subset of 100x100 high density array

For the subset, an 11x11 array was chosen. The attachment pattern was similar to the preliminary 16 sensor array, where the peptides were always placed in the center of the 11x11 array. The peptide locations for the other tests consisted of the next three sensors along the upper right diagonal of the array. Only the upper left octant was depicted(figure 2.9) due to the symmetry of diffusion in all directions.

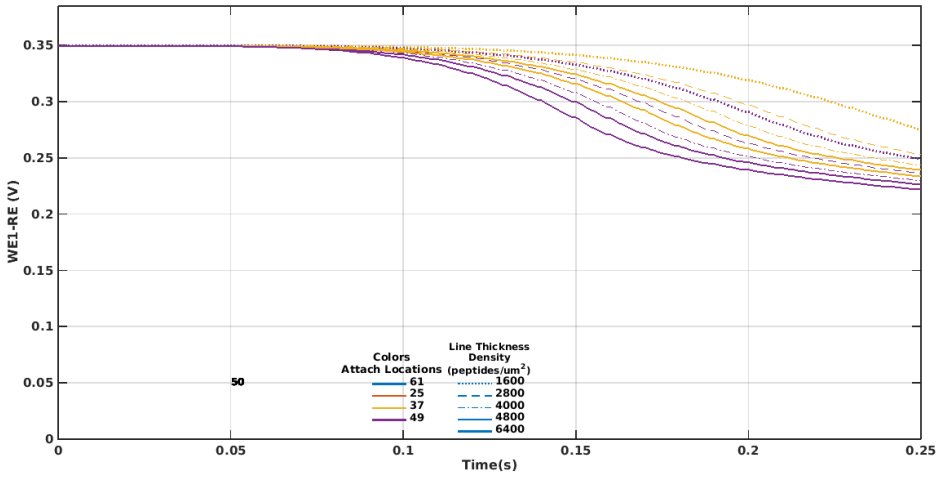
2.4.3 Parameter Variation Crosstalk

In addition to examining crosstalk, two parameters, the sensor capacitance and the peptide potting density, were varied to ascertain their effect on the high density sensor array discharges, especially since the capacitance value was uncertain and required calibration for the preliminary 16 sensor array. The capacitance was varied by orders of magnitude ranging from $10^{-2}pF$ to $10^{3}pF$, while the peptide densities ranged from 1600 to 6400 peptides/μm^2. The capacitance for the high density array will be less than the capacitance of the preliminary array due to the smaller size and a range of smaller capacitances were explored. The proposed peptide potting ranged from 1600 to 6400 peptides/μm^2, which was varied to determine the related reaction rates effects on discharge curves. Crosstalk analysis for the high density subset was performed in the same manner as for the preliminary sensor array. The center sensor, location 61, was active throughout every test and peptides were added in isolated tests to locations 25, 37, and 49, which lie along the diagonal in figure 2.10a.

Each subplot in figure 2.10a and 2.11a represents an individual sensor and depicts a discharge curve for that sensor over a simulated one minute experiment that include reactions and diffusion. The curves in figure 2.10 are simulated using a capacitance of 0.0238pF and varying peptide potting densities. As with the preliminary 16 sensor crosstalk analysis, the discharge curve for sensor 61 is examined to determine the affect of crosstalk from the neighboring sensors along the diagonal. There is no difference in the lines for sensor 61 in figure 2.10a, which indicates the discharge at sensor 61 is not affected by the diffusion of redox molecules from neighboring sensors. The discharge curve is affected by the greater contribution of local redox concentrations, which rapidly discharge the sensors before redox molecules from neighboring areas diffuse across. The variations in density seem to have a minimal affect on the discharge curves, but these difference may be enough to distinguish higher affinity antibody binding from lower



(a) Crosstalk for 121 sensor array while varying peptide potting densities



(b) Zoomed in view of sensor 50 for a closer look at the peptide density effect on crosstalk

Figure 2.10: Crosstalk analysis for 121 sensor subset while varying peptide potting densities

affinity bindings.

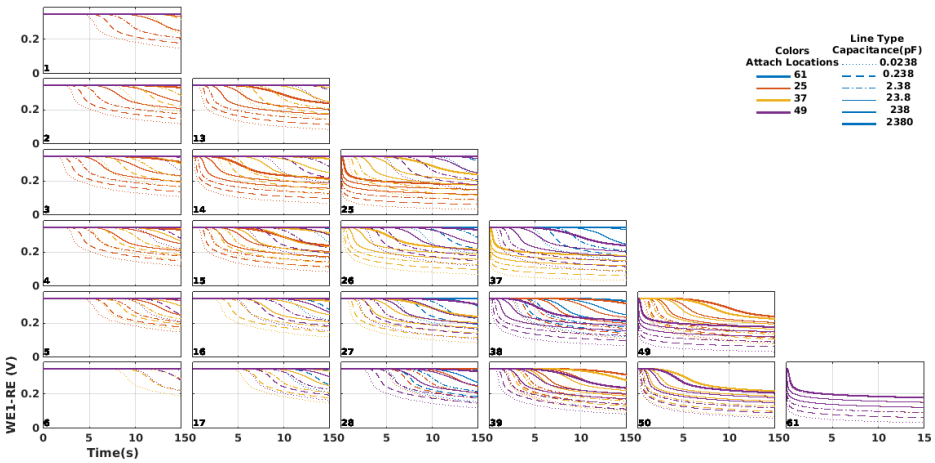
There is, however, a more pronounced effect from crosstalk to the neighboring sensors that do not have any locally produced redox molecules (no antibody binding). In order to observe this effect more closely, the discharge curve for sensor 50 is enlarged and zoomed in time. Figure 2.10b shows the effects of the diffusion from the different testing

locations. The purple curve depicts the test, where redox molecules were diffusing from both adjacent sensors 49 and 61. The other tests have a similar effect on the discharge curves for sensor 50 because sensor 61 is the only adjacent active sensor, which has a much larger effect on the crosstalk due to the shorter diffusion distance needed for the redox molecules. Based on the capacitance of $0.0238pF$ there are about 60 ms before the sensor registers any changes in potential. This means that the time before this occurs contains trustworthy information undistorted by crosstalk noise.

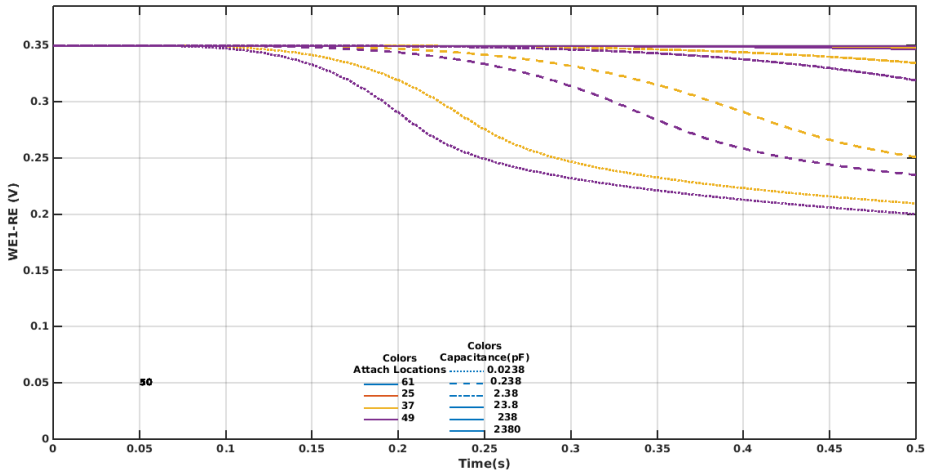
Figure 2.11 showed a similar crosstalk analysis, but instead, the capacitance was varied while maintaining a constant peptide potting density of 1600 peptides/μm^s. Varying the capacitance did not change the results demonstrated by figure 2.10. The active sensor 61 was not affected by the neighboring diffusion of redox molecules. However, the affect of varying capacitance was much more pronounced than the variation of peptide densities. This may be partly attributed to the fact that the capacitance was varied by order of magnitudes whereas the peptide potting density was only varied by a constant amount. The zoomed in view depicted in figure 2.11b, showed that the minimal time for crosstalk to occur was approximately 75ms, but looking back to figure 2.11a, the maximum time needed for crosstalk to show an effect was around 2 to 3 seconds. The longer period can allow for a more stable system and more data measurements to be sampled.

2.4.4 Discussion

The crosstalk and parameter variation provide insight into the signal processing issue. Parameter variation shows that the capacitance of the sensors greatly impacts the sensor discharge curves. Sensors with higher capacitances have a longer period before crosstalk occurs due to more redox molecules needed to shuttle charges between the larger capacitor. The capacitance can be varied by several parameters including the



(a) Crosstalk for 121 sensor array



(b) Zoomed in view of sensor 50 for a closer look at the capacitance density effect on crosstalk

Figure 2.11: Crosstalk analysis for 121 sensor subset while varying sensor capacitance value

temperature of the device, the electrode geometry, and the electrolyte solution. This parameter can altered within physical possibilities to optimize the crosstalk effect. The peptide density variation had less effect, but the variation was less than the capacitor. The actual variation between antibody binding concentrations may be much larger since there can be many antibodies that bind to a single peptide.

Based on the simulations, crosstalk did not effect sensors actively producing the pAP redox molecule because the local concentration was much larger than the slow diffusion effect from neighbors. This is promising as the peptides with large discharge rates are likely to contain more distinct information for a particular disease. However, the simulations did show crosstalk between active sensors and non-active sensors. The redox molecules diffuse from the active sensor to adjacent sensors and cause a discharge if the adjacent sensor does not have locally produced redox molecules. The discharge of the neighboring sensors does not occur immediately since it relies on diffusion.

There are several ways to deal with the crosstalk noise. One method is to only take measurements in the initial period before crosstalk occurs. An issue with this is that the extremely short duration within which this effect occurs may not be long enough for stable readings to be made. In addition, the biological processes may need more time for the reactions to initiate and proceed. A second method for dealing with crosstalk is to use the delay interval of the adjacent sensors as a threshold. This threshold can be used to separate the sensors that began discharging later, which indicates non-locally produced redox molecules due to a longer diffusion time and longer distance. A final method is to introduce redundant peptides and arrange them in spatial patterns that allows signal patterns to be extracted from the deconvolution of the combined signals.

The crosstalk analysis has demonstrated the noise issues that may occur and possible solutions. Although crosstalk occurs, the discharges from the adjecent sensors may not degrade the performance of the disease classifiers. For future works, it will be interesting to see how the end to end simulation, from high density diffusion to classification, performs as the parameters are varied and different signal processing approaches are attempted.

Chapter 2, in part, have been submitted for publication of the material as it may appear in Biomedical Circuits and Systems Conference, 2015, Sun, Alexander; Au,

Anthony; Venkatesh, A.G.; Gilja, Vikash; Hall, Drew A.,IEEE, 2015. The thesis author was a coauthor of this paper.

Chapter 3

Data Analysis: Disease Identification

While optimizing the immunosignaturing chip design is important, the main goal for developing this device is for use as a diagnostic device. For detection of diseases, the information obtained from the IDE sensor discharge readings need to be converted from large arrays of numbers into human interpretable information about each disease state. In order to perform this conversion, machine learning algorithms are used to detect distinct patterns for each class based on the binding to peptides.

3.1 ASU Dataset

The ASU group led by Phillip Stafford and Stephen Johnston, the inventors of immunosignaturing, performed various experiments with immunosignaturing technology. The group used fluorescent imaging techniques to obtain their immunosignature readings, which consist of attaching fluorescent probe to the secondary detection antibody instead of an enzyme for redox reactions [23]. An imaging device was used to scan the microarray with a laser and imaging software assigned fluorescent intensities for each peptide feature in the microarray. Fluorescent intensities measurements by the imaging software were median normalized. In order to test possible classification techniques for the elec-

trochemical immunosignaturing device, public immunosignaturing data was transformed into data relatable to IDE sensor discharge rates. Machine learning algorithms were applied to the newly transformed data for a classification analog to the electrochemical immunosignature device.

The two datasets were collected from many locations across the United States [23]. The first dataset, under accession number GSE52580 in the GEO public repository, contained five different cancer classes and one healthy class for a total of six classes [23]. It had a total of 240 samples with 40 samples per class. The second dataset, under accession number GSE52581 in the GEO public repository, contained 14 different cancer classes and one healthy class for a total of 15 classes [23]. This dataset had a total of 1516 samples, but they were unevenly distributed between classes. The number of samples per class ranged between as few as 5 and as many as 66.

3.1.1 Thresholding

Since the ASU datasets are detected using optical imaging (fluorescence), the exact concentrations of antibodies bound to peptides during the experiments were unknown. The first idea that connects the optical immunosignaturing data to the proposed electrochemical immunosignaturing method is to transform the fluorescent intensity values into relative concentration values. The concentrations are then used as inputs for the sensor discharge simulation. This method will be discussed with more detail in chapter 4. The second idea is to set a threshold that splits the continuous fluorescent intensities into binary values and directly use them to classify the diseases. These binary values ([0,1]) represent peptides that have a low binding affinity to antibodies (below threshold value) and peptides that have a high binding affinity (above threshold). Binding affinities affect the number of antibodies that attached and subsequently, the number of redox molecules produced by the enzymes on the antibodies. As seen in figure 1.4b, higher

concentrations of redox molecules result in higher discharge rates of the IDE sensors. Following this logic, the peptides above the threshold represent high discharge rates and as the threshold is raised, the number of informative peptides becomes more selective. Figure 3.1 shows the fluorescent intensity values obtained for the first dataset. The y-axis represents the different peptide features and the x-axis represents various samples grouped together by class. The image only shows the data within 3 standard deviations of the mean due to very large fluorescent intensity outliers. Figure 3.2 shows the data after the thresholding transformation and feature selection to reduce the dimensionality. The

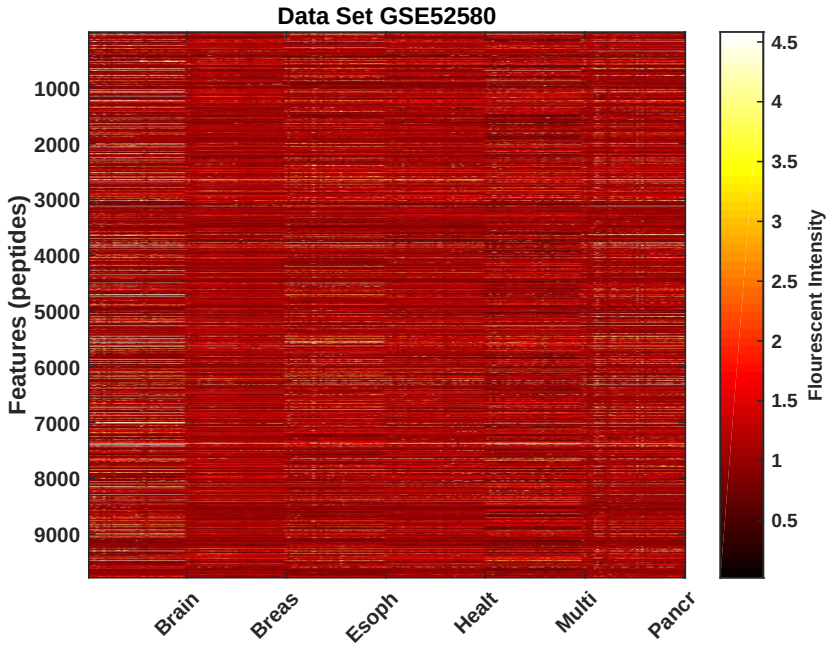


Figure 3.1: GSE52580 dataset immunosignaturing fluorescence intensity data. The x-axis corresponds to samples and the y-axis corresponds to peptide features

threshold value is swept from 0.1 to 4.5 fluorescent intensity units by increments of 0.05. The threshold range is chosen based on the spread of the data. The lower range is set to 0.1, which is close to the 1st percentile of the fluorescent intensities. The upper range is set to 4.5, which is the 99th percentile of the fluorescent intensities. Several outliers have fluorescent intensities a few orders of magnitude greater than the lower 99% of

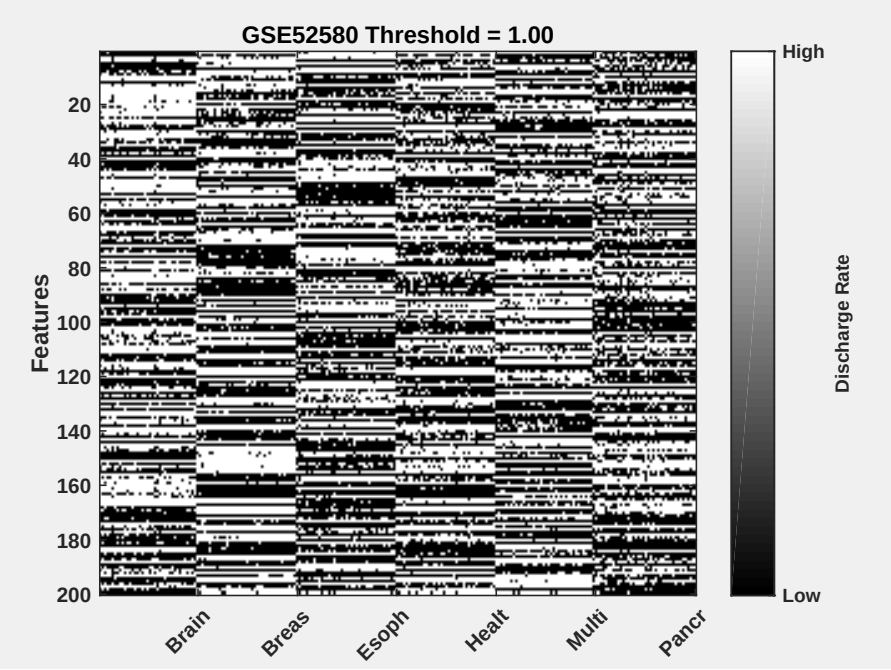


Figure 3.2: GSE52580 with a threshold applied at a fluorescent intensity equal to 1 and top 200 features selected by mRMR

the data, which makes it difficult to sweep through the entire range of intensities. Since the goal of using a threshold is to examine how selectively choosing peptides influences disease detection, the thresholds from 0.1 to 4.5 provides a sufficient range to observe the classification trends.

3.2 Classification

Classification machine learning algorithms are used to detect the various disease states. This approach falls under the category of supervised learning which introduces known labels of each class during the training phase. The other approach falls under the category of unsupervised learning and is known as clustering. This approach lumps data into groups or clusters and predicts which data point belongs to which group. The labels of each data point are not usually known but may be introduced to change to a supervised

learning approach.

3.2.1 Feature Selection

Analysis of Variance (ANOVA) was used as the feature selection technique in the original analysis of the two datasets [23]. However, ANOVA assumes that data is normally distributed, which does not make sense for binary data. Other methods that specifically design for binary and categorical data were needed to handle the thresholded dataset. One method was minimum redundancy maximum relevance (mRMR). This technique, originally designed for DNA microarrays, was computationally efficient and more selective of the features than mutual information or correlation.

3.2.1.1 Minimum Redundancy Maximum Relevance

This feature selection technique was designed for use in DNA microarrays, which have a similar design to immunosignaturing arrays. In both cases the data is very high dimensional with limited samples, making it difficult to implement common machine learning algorithms. The technique, mRMR simultaneously selects features based on two conditions. The first condition, minimum redundancy, aims to minimize the correlation between selected features. Redundant features can cause two main issues during the subsequent classification process. The first issue is computational efficiency [35]. If there are 100 selected features but half of them are highly correlated to other selected features, then computational power is squandered by incorporating the highly correlated features [35]. The second issue that occurs due to redundancy is a lack of generalizability to other datasets. Highly correlated features may only represent a dominant characteristic of the target disease and therefore, the features will be unable to accurately predict the disease if only less common characteristics are present [35]. The second condition, maximum relevancy, aims to maximize the relevance of each feature to the classes. This

selects the features that contain the most useful information to distinguishing each class.

Following these two conditions, Ding and Peng in [35] derived the equations and constraints necessary for the feature optimization problem. First, the minimum redundancy equation finds the features that are most dissimilar from current features in the selected subset.

$$minW_I, \qquad W_I = \frac{1}{|S|^2} \sum_{i,j \in S} f(i,j) \tag{3.1}$$

where S represents the subset of features to be selected, $|S|$ represents the number of features in S, f(i,j) represent any criterion function for evaluating the similarity of the features. The criterion function can be either discrete or continuous. For example, a continuous criterion function is Pearson correlation, while a discrete function is mutual information. Equation 3.1 finds minimum the mean criterion function value evaluated between all combinations of features in the subset S.

$$maxV_I, \qquad V_I = \frac{1}{|S|} \sum_{i \in S} f(h,i) \tag{3.2}$$

where h represents the classes. Equation 3.2 finds the maximum mean criterion function value of features in S evaluated against each class.

The two conditions modeled in equations 3.1 and 3.2 can be combined into a single optimization function with two constraints. The first constraint uses a difference condition to combine the two conditions as seen in equation 3.3a. The second constraint uses a quotient condition to combine the two equations as seen in equation 3.3b.

$$max(V_I - W_I) \tag{3.3a}$$

$$max(V_I / W_I) \tag{3.3b}$$

3.2.2 Algorithms

The following classification algorithms are commonly used and simple to implement. They were used to classify the two optical datasets with highest average accuracies of 95% and 98% respectively [23]. These accuracies reflect the best average performance selected between the different learning algorithms. All classifications were performed with a 4 fold cross validation using 200 features selected by mRMR with the quotient constraint.

3.2.2.1 Naive Bayes Classifier (NB)

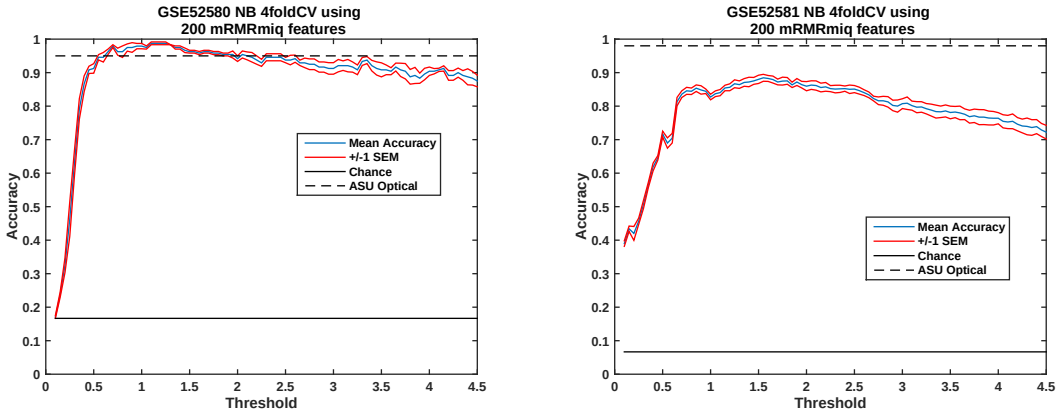
The naive Bayes is one of the simplest machine learning algorithms because it makes the assumption that the features are conditionally independent of each other given the class [36]. This assumption leads to a class conditional probability of the form [36]:

$$p(X|y=c,\theta)=\prod_{j=1}^{D} p(x_j|y=c,\theta_{jc}) \tag{3.4}$$

where c are the different number of classes, D is the total number of features, and θ are the parameters for the model. This assumption of independence given the class is often called naive because features are generally not independent in real life. For instance, it is likely that binding to one peptide will be related to the binding of another peptide due to the physical characteristics of the disease. Although this naive assumption is simple and untrue in the real world, naive Bayes classifiers perform well on many classification tasks (Domingos and Paxxani 1997). One reason for its versatility is the low number of parameters (O(CD), where C is the number of classes and D is the number of features) needed to fit the model , which makes it less susceptible to overfitting [36].

Gaussian distributions are typically used to model the class conditional probabilities for continuous data, but they are more likely to fail when considering binary

data. This is due to the fact that the binary thresholding performed on the data causes various features across a class to be equal to the same value (either 0 or 1). Hence, it results in the Gaussian model for that class conditional probability to have zero variance, which makes it non-ideal to use a Gaussian distribution. However, other probability distributions such as the multinomial distribution are specifically designed to model binary and discrete data. Using the multinomial distribution with naive Bayes is known as the bag of tokens model, commonly used in document classification [36]. This models the count of various words within a document and fits these counts to a multinomial model [36]. For immunosignaturing, the various peptides represent the words and each sample is a document. The classification algorithm was implemented using MATLAB 2014b [37] with the distribution set to 'mvmn' (multivariate multinomial), which assumes each predictor (counts for peptides) follows a multinomial model within a class.



(a) GSE52580 dataset classification accuracy, chance is 1 out of 6 classes

(b) GSE52581 dataset classification accuracy, chance is 1 out of 15 classes

Figure 3.3: Naive Bayes classification through various binary thresholds to simulate low and high discharge rates

Compared to the results of the ASU group, the NB classification for the first dataset performed around the same accuracy. In fact, between a threshold of 0.5 and 2, the classification of the thresholded data performed better than the ASU dataset. The performance on the thresholded second dataset was lower than the average accuracy

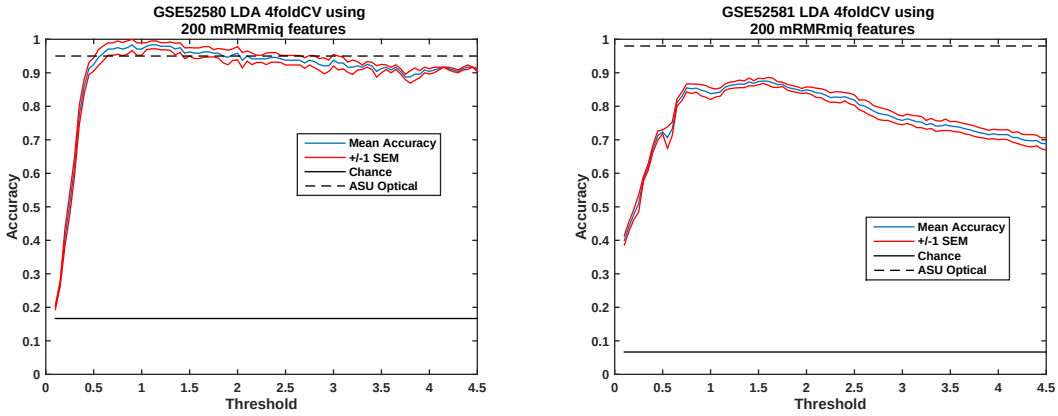
obtained by the ASU optical group.

3.2.2.2 Linear Discriminant Analysis (LDA)

Although using Gaussian distributions to model binary data may not be best choice, LDA is used as a comparison with the methods conducted in the original analysis. Gaussian discriminant analysis, which LDA falls under, calculates a full covariance matrix of all the features. This results in a multivariate Gaussian distribution, with the parameters of mean and covariance, modeled for each class. However, if the diagonal of the covariances for each class are used instead of the full covariance matrix (assumes features are uncorrelated), then the model becomes equivalent to the Naive bayes classifier[36]. Linear discriminant analysis assumes that the covariances are equal for each class, which creates linear separating boundaries between classes. Quadratic discriminant analysis (QDA) does not make this assumption of equal class covariance, which allows for non-linear boundaries[36], but QDA may result in overfitting if there are few training samples due to non-linear boundaries created for outliers[38]. This makes QDA an unwise choice for this sample limited dataset.

For the binary data transformation, LDA ran into similar problems as Naive bayes because of cases where there was zero variance for a feature across a class. In order to correct for this issue, the diagonal of the covariance matrix was used instead of the full matrix. This modification, called diagonal LDA, differs from Naive bayes because of the additional constraint that makes the covariances of all classes equal. In high dimensional problems, this model can work better than a full covariance LDA model [39]. Additionally, in cases of a singular matrix, the MATLAB implementation used the pseudoinverse of the covariance when appropriate[37].

Compared to the results of the ASU group, the LDA classification for the first dataset performed around the same accuracy. Between a threshold of 0.5 and 1.5,



(a) GSE52580 dataset classification accuracy, chance is 1 out of 6 classes

(b) GSE52581 dataset classification accuracy, chance is 1 out of 15 classes

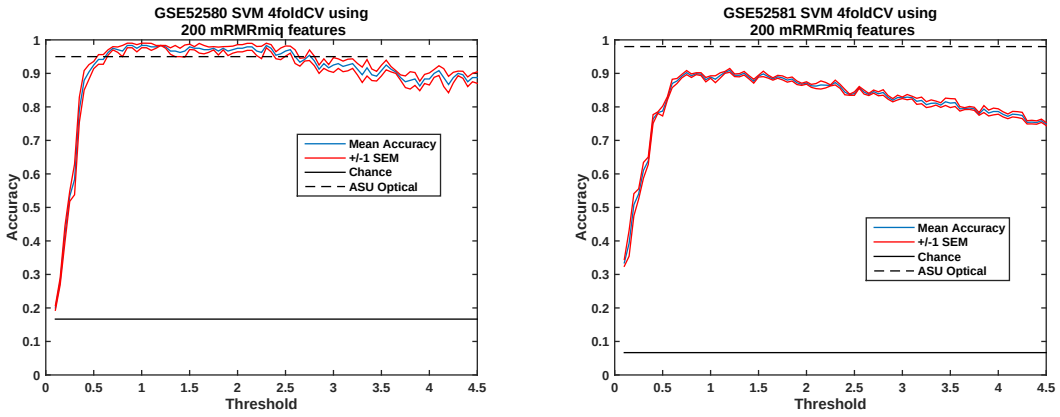
Figure 3.4: LDA classification through various binary thresholds to simulate low and high discharge rates

the classification of the thresholded data performed better than the ASU dataset. The performance on the thresholded second dataset was lower than the average accuracy obtained by the ASU optical group.

3.2.2.3 Support Vector Machines (SVM)

SVM constructs a hyperplane or set of hyperplanes that separates the classes, but do not use probability distributions to model the data. Instead, SVMs only focus on the points near the separating boundary that are most difficult to separate due to overlap between the classes. In this regard, the SVM does not model the structure of the data and merely aims for classification performance, which falls under the category of discriminative classifiers [40]. The SVM uses support vectors to calculate the separating hyperplane(s) by relying on support vectors. These support vectors consists of the data points closest to the separating boundary drawn as parallel vectors to the separating hyperplane vector [41]. SVMs can also take advantage of soft margins, which allows some data points to fall on the incorrect side of the hyperplane [38]. The size of the margins between the support vectors can be varied, which leads to a bias-variance tradeoff

due to the changing number of data points used in the support vectors and tolerance for error. The kernel function of an svm decides whether the boundaries are linear or non-linear, which may be appropriate depending on the type of problem. For the thresholded data a linear kernel function was used, since the non-linear boundary can lead to overfitting due to the limited number of samples. Also, typically SVMs are constructed for problems containing binary classes (two classes) due to the concept of separating hyperplanes. In order to extend SVMs for a multiclass problem, a one versus all approach is used. One versus all consists of performing K (total number of classes) binary classifications, where one class is considered and all other classes are grouped into a single class [38]. This results in classification only between classes and the process is repeated for the number of classes. The classifications are chosen based on the class that had the largest positive class scores, which indicates how strongly the test sample is predicted to be in that class [37].



(a) GSE52580 dataset classification accuracy, chance is 1 out of 6 classes

(b) GSE52581 dataset classification accuracy, chance is 1 out of 15 classes

Figure 3.5: SVM classification through various binary thresholds to simulate low and high discharge rates

Compared to the results of the ASU group, the SVM classification for the first dataset performed around the same accuracy. Between a threshold of 0.5 and 2.5, the classification of the thresholded data performed better than the ASU dataset. The

performance on the thresholded second dataset was lower than the average accuracy obtained by the ASU optical group.

3.3 Clustering

Examination of the data clusters provided more insights into the structure of the data as well as possible clustering approaches to explore. The thresholded optical data with the top 200 features selected through mRMR with the quotient constraint was used for the initial visualization purpose. A threshold of one was selected, since this value corresponded to high performance from all of the classifiers. In order to visualize the high dimensional data, the features axes were rotated and projected onto the top three principal components, which consist of the three vectors with the highest data variance.

The visualization of first dataset with six classes is shown in figure 3.6. Based on this projection, it can be seen that the 6 classes are very well separated into tightly packed individual clusters. The high classification accuracy can be attributed to these clean separations. The visualization for the second dataset, 3.7, is harder to discern due to the larger number of classes. This does not necessarily mean that the classes in the second dataset are not well separated, since three dimensions may not be enough to properly visualize the high dimensional data.

This is only a preliminary look at the clusters, but the evidence of these well separated classes shows the potential for possible clustering approaches. An advantage to this approach is that the number of clusters detected by the algorithm can be varied. This allows more flexibility in determining the structure of the data because there can be cases where patients have multiple diseases at the same time. The combination of the multiple diseases could produce a different immunosignature pattern than simply the combination of the patterns for each individual disease, which can appear as a separate distinct cluster

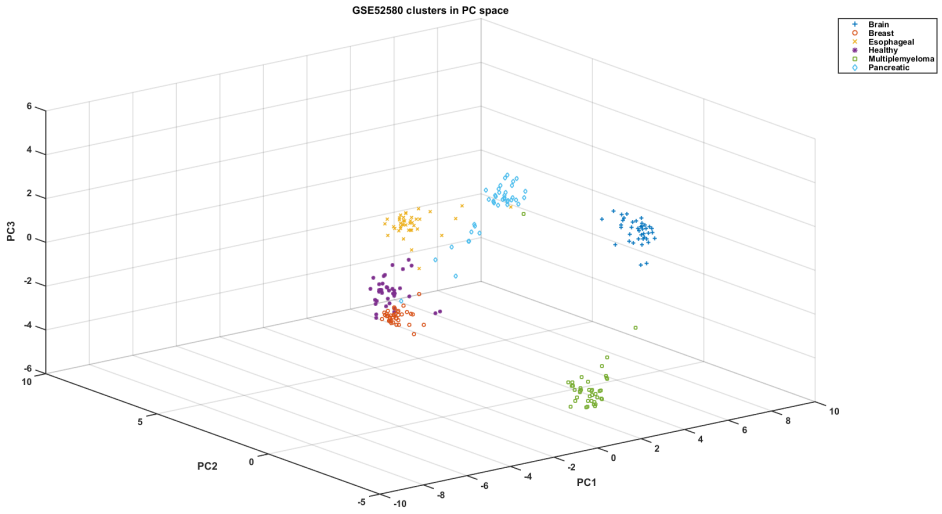


Figure 3.6: GSE52580 cluster visualization within PC space

from the clusters specific to a single disease.

3.4 Discussion

The results for the classification of the thresholded data demonstrated that good performance can be achieved even with the loss of information due to the conversion to binary data. In fact, the performance was better than the performance obtained by the ASU group using the continuous optical data for first dataset. One reason may be that the thresholding filtered out some noise contained within the readings with lower intensity values. Many types of noise can be introduced during the fluorescent imaging process including background fluorescence, magnification of the fluorescent signal, the specimen itself, the scanner and imaging software, and preparation/handling of the samples [42]. Another reason for the improved performance may be attributed to the feature selection technique. The ASU optical group used the technique called ANOVA, which finds the features that most distinguish between different classes. As discussed in

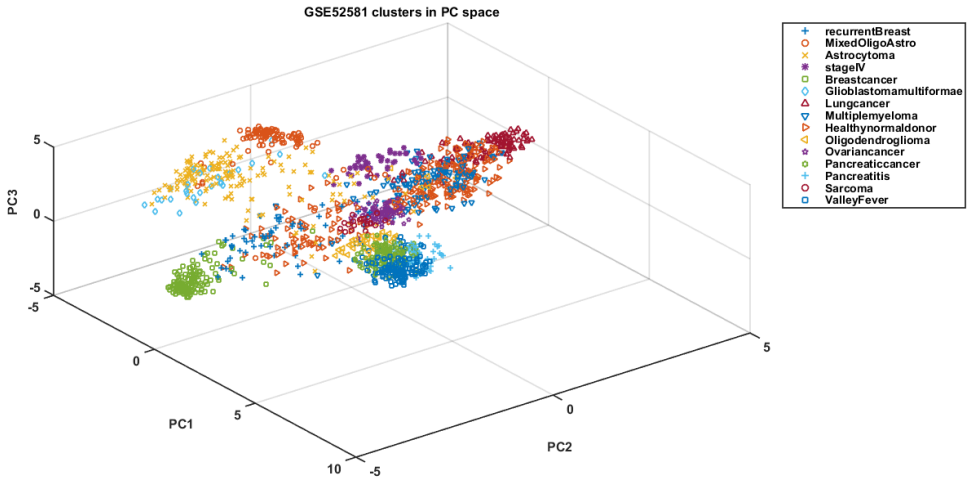


Figure 3.7: GSE52580 cluster visualization within PC space

section 3.2.1.1, mRMR finds features with both the least repeated information and the most class distinguishing ability. The combination filtering of noise and the selection of more relevant features produced better performance for the first dataset, but the performance for the second dataset was less than the performance obtained by the ASU group. This may be attributed to the fact that important distinguishing information was contained in the lower intensity features, which were filtered out by the thresholding approach.

The classification performance of both datasets also demonstrate stability as the thresholds increased and the number of high discharge rate peptides became more selective. The result indicates the presence of a few very distinguishing peptides for these diseases with potentially high discharge rates of electrochemical sensors. Further inspection of these peptides could lead to insights on the disease pattern structure and correlated production of different antibodies. This knowledge can be used for future exploration of a device containing a much smaller array of sensors using the identified very informative peptides. This smaller device will have a trade-off between a cheaper, less noisy implementation and a loss of generalizability due to overfitting for specific

diseases. Once the high density device is fabricated, true electrochemical data can be obtained to explore this possibility.

The machine learning approaches were chosen for comparison to the ASU analyses. Specific approaches were not optimized, allowing for a wide variety of improvements to performance, ranging from choosing more complex approaches that better model the data structure to optimizing the parameters for NB, LDA, and SVM. More complicated approaches include temporal models such as the hidden Markov model, which are able to leverage the time sequenced data obtained from measuring the discharge curves over time. In addition to classification, clustering approaches such as Gaussian mixture models, latent Dirichlet allocation (discrete), and mixed-membership naive Bayes model allow one to uncover possible clusters or classes that consist of the combination of multiple diseases. Last of all, experimenting with different feature selection techniques can lead to better performance and help indicate important peptides for certain diseases. These preliminary investigations into the electrochemical classification as well as the ASU work on the optical data show the potential of immunosignatures and have many avenues for improvements.

Chapter 4

From Simulated to Real

This chapter aims to assimilate the knowledge from chapters 2 and 3 into a mock end to end simulation, beginning from the redox concentrations, to sensor discharge readings, and finally disease diagnosis. This is only a preliminary examination of the potential device effectiveness by linking the real data to the simulated data, the caveat being that there is no exact method for transforming the fluorescent intensities from the optical imaging to concentration values. For the preliminary analysis, a simple linear mapping was used between the fluorescent intensities and concentration of antibodies bound in a particular well. This approach does not necessarily quantitatively translate to the electrochemical readings, but at least the relative fluorescent intensities should reflect relative concentrations(i.e. higher fluorescent intensities reflects more binding, while lower fluorescent intensities reflect less binding).

4.1 Method:Scaling Fluorescent Data

This approach was designed to explore the possibilities of the electrochemical sensing in a simple manner in order to gain some basic intuition. A linear relationship was derived between the optical fluorescent intensity value and the concentration of

redox molecules. To determine this linear relationship, an intensity value was chosen as the maximum intensity. This value was mapped to the maximum concentration that was determined by running the diffusion simulation of the 10,000 peptide array for five minutes. This interval was chosen somewhat arbitrarily since the exact experimental duration time for the optical data was not known. The five minutes was longer to accomodate the extra time needed for the optical detection procedures but not too long due to computation time. The maximum intensity relating to the redox concentration was not chosen to simply be the maximum optical intensity value because the data contained many outliers. The histogram of the first optical dataset showed the majority of the fluorescent intensities had values between 0 and 12, but many outliers had much larger intensities ranging to 350. The intensity value for the maximum was chosen to be the point where the long tail of the histogram began. The second point for the linear relationship, related the lowest intensity value to a redox concentration of zero. For the initial mapping, a value of zero intensity was chosen, which also meant the intercept was zero. The slope of the line connecting these two points was used as the linear

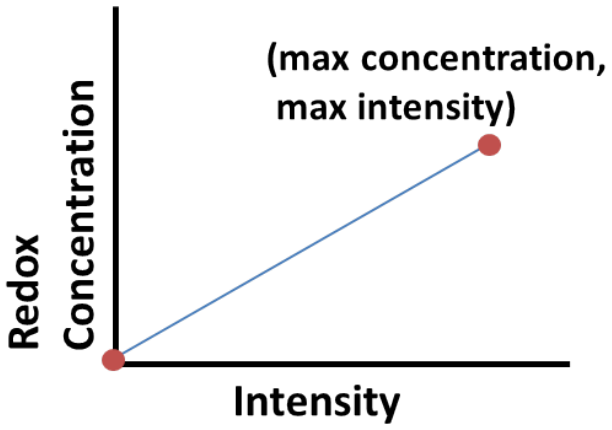


Figure 4.1: Illustration of linear relationship between redox molecule concentration and fluorescence intensity

scaling. Ideally, a range of values for the maximum and minimum intensities would be swept through to determine the optimal value. Once the initial scaling was obtained, the

transformed concentration values were used as an input for the sensor discharge model. The discharges were simulated for 0.5 seconds, since only the initial slopes were used as features and not the whole time course data. The simulation also used a capacitance value of $0.0238pF$.

The initial slopes were extracted from the discharge data by taking the difference over the interval 0-20ms. This short interval was chosen to get the immediate slope and avoid crosstalk of the sensors. Even though no crosstalk occurs for the optical data, the parameters were chosen as if there were in order to more closely parallel the electrochemical sensing paradigm. Once the features were extracted, the continuous analog values were quantized into integers. This step simulated the analog to digital converter(ADC) used to obtain the digital information. A 1.8V, 12 bit ADC was proposed for use in the high density array, which corresponds to a least significant bit 0.44mV. The ADC was simulated because it can introduce noise by cutting off information contained in decimal places smaller than $0.44x10^{-}3$.

The last step was to classify the simulated discharge data. The same three classification algorithms were used as chapter 3. Since the dischrage data was no longer binary, gaussian distribution models were used for naive Bayes and LDA. In addition, LDA was able to perform with the full covariance matrix instead of only the diagonal. SVM still used a linear kernel with one vs all classification. Classifications were performed with all of the features and with feature selection. For all of the features, actually several of the peptide features were removed in the second dataset due to NaN values for an entire class. Having these present resulted in zero within class variance leading to failed Gaussian distributions for NB and LDA. For feature selection, mRMR with the quotient constraint and mutual information as the criterion function, was used to reduce the dimension of the simulated discharge data down to 200 peptides. This method makes many assumptions in order to infer the scaling relation between the

optical data and redox concentration values. This model contains a caveat other than the assumptions made to infer the linear transformation terms. The discharge slopes reflect the relative concentrations at a single point in time. because the values from the optical data only show the final time point of antibody binding. The time evolution of the redox concentrations is lost, which means that this conversion of optical to electrochemical simulates a local discharge at that time point. This simulation is equivalent to taking a slice in time of the discharge curves.

4.2 Results:Discharge Readings

Table 4.1 shows the classification performances when all of the peptides features are used. Four fold cross-validation was performed on all three of the algorithms. The accuracy presented in the table results from the average of the four fold cross-validation. Naive Bayes had the best performance on the first dataset but not significantly different than the other two classifiers. SVM performed the best on the second dataset, much better than NB. The data may not be normal, which made using a Gaussian probability distribution not ideal to try and model the structure of the data. The accuracies for

Table 4.1: Classification performance using all features

	No ADC Noise		**ADC Noise**	
	Dataset1	**Dataset2**	**Dataset1**	**Dataset2**
NB Accuracy	0.9667	0.8502	0.9667	0.8502
SVM Accuracy	0.9542	0.9519	0.9542	0.9525
LDA Accuracy	0.9625	0.9202	0.9625	0.9255
ASU Optical	*0.95*	*0.98*		
Thresholded Data	*0.9875*	*0.909*		

first data set were slightly higher than the ASU optical data but slightly lower than the

thresholded data. The performance on the second dataset was lower than ASU optical but higher than the thresholded data. It seems that thresholding was better for the less complicated dataset with only six classes, but the continuous data proved better for the larger dataset with 15 classes. Also, the ADC noise did not significantly affect the classification performance.

Table 4.2: Classification performance using 200 mRMR features

	No ADC Noise		**ADC Noise**	
	Dataset1	**Dataset2**	**Dataset1**	**Dataset2**
NB Accuracy	0.9750	0.7897	0.9792	0.7897
SVM Accuracy	0.9667	0.9123	0.9625	0.9130
LDA Accuracy	0.8458	0.8938	0.8500	0.8951
ASU Optical	*0.95*	*0.98*		
Thresholded Data	*0.9875*	*0.909*		

The number of features were reduced to 200 in order to remove noise and reduce computation. The resulting classification performances shown in table 4.2 were calculated with four fold cross-validation and the first 200 features chosen by mRMR with the quotient constraint. Same as all the features, NB performed the best on the first dataset and SVM performed the best on the second dataset. The performance was slightly raised by feature selection for the first dataset, but it was lowered for the second dataset. These accuracies were comparable to the thresholded data. The 12 bit ADC noise did not significantly affect the performance.

4.3 Discussion

This initial look at the a full end to end simulation, albeit with several caveats, shows the potential of the electrochemical device. The data obtained from the actual

device will have issues such as crosstalk noise, but the amplification afforded by the continuous production of redox molecules can produce more separated features. The real electrochemical data will also have the advantage of time evolution, which can allow the use of temporal classification models.

As this model only served as a quick look at a simulated end to end performance, the ideal parameters or methods can be searched for the best performance. The values used for the linear scaling between the intensity and concentration can be varied to change the proportionality. If the data was continuous, then the relative scalings would remain the same and no effect will be seen, but the data was quantized by the ADC, thus scalings resulting in smaller differences will effect the signal distinctness.

As in chapter 3, the classification has much room for improvement, ranging from the specific algorithm to the feature selection. The algorithms were chosen to mirror the same set used by the ASU group, but other algorithms that take advantage of specific aspects of immunosignature structure can have better performance. As mention previously, latent and temporal models will serve to better examine the real electrochemical data. But not only can the choice of algorithm be improved, the tuning parameters can be optimized (e.g. the margin size for the SVM classifier). Feature selection remains an important component and much research has been done in this field relating to similar high dimensional microarrays and gene sequencing. Improvements can results from more complex selection techniques or simply iterating through the optimize the selected number of features.

Finally, the physical device and recording mechanisms have room for improvement. The capacitance is selected as $0.0238pF$, but until a more exact measurement or approximation is obtained, it can varied to examine the effect on overall disease classification performance. In addition, the 12 bit ADC demonstrates little loss of information. Thus the current ADC can be reduced in bit size until the degradation in performance

becomes noticeable. This has the potential to lower costs if a smaller cheaper ADC can be used. The final issue to be aware of is the switching time noise introduced the be mux. The proposed implementation will require 100ns, including settling time, to switch to the next sensor. It will take 1ms to switch through an array of 10,000 sensors, meaning the lowest resolution for each sensor is 1ms. The delay may have an effect on disease performance and will have to be accounted for in the model. The current real to simulated model of the electrochemical device has many areas for improvement, but showcases the exciting future potential of the high density electrochemical biosensor for immunosignatures.

Chapter 5

Conclusion

The development of a high density electrochemical immunosignature biosensor array requires the optimization of many parameters ranging from specific peptide locations to biosensor chip geometry to the particular machine learning model for disease diagnosis. It is costly and time consuming to test each parameter with the physical device. Software simulation, on the other hand, does not incur the costs of fabricating chips with different sensor sizes and avoids the preparation time necessary for conducting experiments.

This work developed the framework for a simulation predictive of the electrochemical immunosignature biosensor array. The simulation was split into two parts consisting of the device physics(diffusion and discharge models) and disease identification. For the first part, validation of the simulated sensor discharge model showed that the real sensor measurements could be accurately modeled given a calibration factor. Once reliability of the model was established, basic crosstalk patterns were analyzed. These patterns revealed little to no crosstalk effect on the sensors actively producing redox molecules but larger effects on adjacent sensors with no local redox production. The results of the crosstalk simulation suggests possible implementation strategies leveraging

the discharge delay due to the speed of diffusion.

Classification approaches demonstrated the richness and robustness of the immunosignaturing data. The three classification algorithms applied were commonly used approaches and not chosen due to particular structure in the data. Even though optimal algorithms were not chosen, the classification accuracies still reached into the 90% range. Moreover, thresholding into binary data and selecting only a few of those features did not result in sizable degradation to accuracy even with the considerable loss of information. The final simulation also contained a loss of information due to the ADC quantization, but performance was unaffected. Preliminary classification highlighted the robustness of immunosignature data, which can tolerate various sources of noise while maintaining high accuracy. Optimizing these approaches could lead to near perfect disease detection and discovery of the structure within the data.

5.1 Future Work

With the simulation ground layer built, the natural next step is to validate the whole simulation. The diffusion of redox molecules needs to be verified, since all other parts of the simulation depend on the subsequent concentration values. Once the simulation has been validated on the preliminary 16 sensors array and calibrated for the high density sensor, it can be extended to the optimize the high density array parameters for the best signal to noise ratio.

A great deal of room exists for the improvement of the signal processing to classification pipeline. For instance, optimal parameters can be found by sweeping through a host of values. Applying different classification approaches, such as temporal models that leverage the time course information from the discharge curves, may lead to improved disease diagnosis. Recent attention in the machine learning field has focused on deep

learning, which consists of neural networks. Deep learning requires numerous samples, which can be obtained if the flu tracker vision is realized. All of the immunosignature data obtained across the nation will be entered into the deep learning architecture to discover the many features and diseases present. Many possibilities exist to improve the current design, but immunosignatures continue to be an abundance of information and potential.

Bibliography

[1] J. B. Legutki, Z.-G. Zhao, M. Greving, N. Woodbury, S. A. Johnston, and P. Stafford, "Scalable high-density peptide arrays for comprehensive health monitoring," *Nat Commun*, vol. 5, Sep 2014, article. [Online]. Available: http://dx.doi.org/10.1038/ncomms5785

[2] CDC. (2015) Weekly us map:influenza summary update. [Online]. Available: http://www.cdc.gov/flu/weekly/usmap.htm

[3] Google. (2015) Google flu trends data. [Online]. Available: www.google.org/flutrends/about/

[4] J. C. Davis, L. Furstenthal, A. a. Desai, T. Norris, S. Sutaria, E. Fleming, and P. Ma, "The microeconomics of personalized medicine: today's challenge and tomorrow's promise." *Nature reviews. Drug discovery*, vol. 8, no. 4, pp. 279–286, 2009.

[5] M. S. Penn, M. A. Yenikomshian, A. K. G. Cummings, A. Klemes, J. M. Damron, S. Purvis, M. Beidelschies, and H. G. Birnbaum, "The economic impact of implementing a multiple inflammatory biomarker-based approach to identify, treat, and reduce cardiovascular risk," *Journal of Medical Economics*, vol. 18, no. 7, pp. 483–491, 2015, pMID: 25763924. [Online]. Available: http://dx.doi.org/10.3111/13696998.2015.1029490

[6] G. Poste, "Bring on the biomarkers." *Nature*, vol. 469, no. 7329, pp. 156–157, 2011.

[7] P. Stafford, R. Halperin, J. B. Legutki, D. M. Magee, J. Galgiani, and S. A. Johnston, "Physical characterization of the "immunosignaturing effect"," *Mol. Cell Proteomics*, vol. 11, 2012.

[8] N. L. Anderson and N. G. Anderson, "The human plasma proteome: History, character, and diagnostic prospects," *Mol. Cell Proteomics*, vol. 2, no. 1, p. 50, 2003. [Online]. Available: http://www.mcponline.org/content/2/1/50.short

[9] S. S. Hori and S. S. Gambhir, "Mathematical model identifies blood biomarker based early cancer detection strategies and limitations," *Science Translational Medicine*, vol. 3, no. 109, pp. 109ra116–109ra116, 2011.

[10] P. O. Brown and C. Palmer, "The preclinical natural history of serous ovarian cancer: Defining the target for early detection," *PLoS Med*, vol. 6, no. 7, p. e1000114, 07 2009.

[11] B. Sulzer, J. L. van Hemmen, A. U. Neumann, and U. Behn, "Memory in idiotypic networks due to competition between proliferation and differentiation," *Bulletin of Mathematical Biology*, vol. 55, no. 6, pp. 1133 – 1182, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092824005801685

[12] S. Cenci and R. Sitia, "Managing and exploiting stress in the antibody factory," *{FEBS} Letters*, vol. 581, no. 19, pp. 3652 – 3657, 2007, cellular Stress. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0014579307004218

[13] J. Cooperman, R. Neely, D. T. Teachey, S. Grupp, and J. K. Choi, "Cell division rates of primary human precursor b cells in culture reflect in vivo rates," *STEM CELLS*, vol. 22, no. 6, pp. 1111–1120, 2004. [Online]. Available: http://dx.doi.org/10.1634/stemcells.22-6-1111

[14] I. FÃűrster and K. Rajewsky, "The bulk of the peripheral b-cell pool in mice is stable and not rapidly renewed from the bone marrow." *Proceedings of the National Academy of Sciences*, vol. 87, no. 12, pp. 4781–4784, 1990. [Online]. Available: http://www.pnas.org/content/87/12/4781.abstract

[15] Z. Hao and K. Rajewsky, "Homeostasis of peripheral b cells in the absence of b cell influx from the bone marrow," *The Journal of Experimental Medicine*, vol. 194, no. 8, pp. 1151–1164, 2001. [Online]. Available: http://jem.rupress.org/content/194/8/1151.abstract

[16] G. P. Dunn, L. J. Old, and R. D. Schreiber, "The immunobiology of cancer immunosurveillance and immunoediting," *Immunity*, vol. 21, no. 2, pp. 137 – 148, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1074761304002092

[17] Y. Kotera, J. D. Fontenot, G. Pecher, R. S. Metzgar, and O. J. Finn, "Humoral immunity against a tandem repeat epitope of human mucin muc-1 in sera from breast, pancreatic, and colon cancer patients," *Cancer Research*, vol. 54, no. 11, pp. 2856–2860, 1994.

[18] E. Jäger, Y. T. Chen, J. W. Drijfhout, J. Karbach, M. Ringhoffer, D. Jäger, M. Arand, H. Wada, Y. Noguchi, E. Stockert, L. J. Old, and a. Knuth, "Simultaneous humoral and cellular immune response against cancer-testis antigen ny-eso-1: definition of human histocompatibility leukocyte antigen (hla)-a2-binding peptide epitopes." *The Journal of experimental medicine*, vol. 187, no. 2, pp. 265–270, 1998.

[19] J. M. Reiman, M. Kmieciak, M. H. Manjili, and K. L. Knutson, "Tumor immunoediting and immunosculpting pathways to cancer progression," *Seminars in Cancer Biology*, vol. 17, no. 4, pp. 275 – 287, 2007, making the Tumor-Specific Effectors Ineffective. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1044579X0700034X

[20] M. E. Hudson, I. Pozdnyakova, K. Haines, G. Mor, and M. Snyder, "Identification of differentially expressed proteins in ovarian cancer using high-density protein microarrays," *Proceedings of the National Academy of Sciences*, vol. 104, no. 44, pp. 17 494–17 499, 2007. [Online]. Available: http://www.pnas.org/content/104/44/17494.abstract

[21] R. B. Darnell and J. B. Posner, "Paraneoplastic syndromes involving the nervous system," *New England Journal of Medicine*, vol. 349, no. 16, pp. 1543–1554, 2003, pMID: 14561798. [Online]. Available: http://dx.doi.org/10.1056/NEJMra023009

[22] J. B. Legutki and S. A. Johnston, "Immunosignatures can predict vaccine efficacy," *Proceedings of the National Academy of Sciences*, vol. 110, no. 46, pp. 18 614–18 619, 2013. [Online]. Available: http://www.pnas.org/content/110/46/18614.abstract

[23] P. Stafford, Z. Cichacz, N. W. Woodbury, and S. A. Johnston, "Immunosignature system for diagnosis of cancer," *Proceedings of the National Academy of Sciences*, vol. 111, no. 30, pp. E3072–E3080, 2014. [Online]. Available: http://www.pnas.org/content/111/30/E3072.abstract

[24] B. O. Donnell, A. Maurer, A. Papandreou-suppappola, and P. Stafford, "Time-Frequency Analysis of Peptide Microarray Data : Application to Brain Cancer Immunosignatures," vol. 14, pp. 219–233, 2015.

[25] J. R. Brown, P. Stafford, S. a. Johnston, and V. Dinu, "Statistical methods for analyzing immunosignatures," *BMC Bioinformatics*, vol. 12, no. 1, p. 349, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/349

[26] M. Kukreja, S. Johnston, and P. Stafford, "Comparative study of classification algorithms for immunosignaturing data," *BMC Bioinformatics*, vol. 13, no. 1, 2012. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-13-139

[27] P. Delahay, "Coulostatic method for the kinetic study of fast electrode processes. i. theory," *The Journal of Physical Chemistry*, vol. 66, no. 11, pp. 2204–2207, 1962. [Online]. Available: http://dx.doi.org/10.1021/j100817a030

[28] W. H. Reinmuth and C. E. Wilson, "An impulse (coulostatic) relaxation method for the study of rapid electrode processes." *Analytical Chemistry*, vol. 34, no. 9, pp. 1159–1161, 1962. [Online]. Available: http://dx.doi.org/10.1021/ac60189a002

[29] A. J. Bard and L. R. Faulkner, *Electrochemical methods : fundamentals and applications*, 2001.

[30] D. G. Sanderson and L. B. Anderson, "Filar electrodes: steady-state currents and spectroelectrochemistry at twin interdigitated electrodes," *Analytical Chemistry*, vol. 57, no. 12, pp. 2388–2393, 1985. [Online]. Available: http://dx.doi.org/10.1021/ac00289a050

[31] A. J. Bard, J. A. Crayston, G. P. Kittlesen, T. V. Shea, and M. S. Wrighton, "Digital simulation of the measured electrochemical response of reversible redox couples at microelectrode arrays: consequences arising from closely spaced ultramicroelectrodes," *Analytical Chemistry*, vol. 58, no. 11, pp. 2321–2331, 1986. [Online]. Available: http://dx.doi.org/10.1021/ac00124a045

[32] X. Zhu, J.-W. Choi, and C. H. Ahn, "A new dynamic electrochemical transduction mechanism for interdigitated array microelectrodes," *Lab Chip*, vol. 4, pp. 581–587, 2004. [Online]. Available: http://dx.doi.org/10.1039/B407930B

[33] K. F. Sykes, J. B. Legutki, and P. Stafford, "Immunosignaturing: a critical review," *Trends in Biotechnology*, vol. 31, no. 1, pp. 45 – 51, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167779912001928

[34] W. A. Ritschel, *Handbook of basic pharmacokinetics*, 1980.

[35] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, vol. 3, no. 2, pp. 185–205, 2003.

[36] K. P. Murphy, *Machine Learning a Probabilistic Perspective*, 1st ed. Cambridge Massachusetts: The MIT Press, August 2012.

[37] MATLAB, *version 8.4.0.150421 (R2014b)*. Natick, Massachusetts: The MathWorks Inc., 2014.

[38] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction toStatistical Learning*, 1st ed. New York: Springer-Verlag New York, 2013.

[39] P. J. Bickel and E. Levina, "Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, no. 6, pp. pp. 989–1010, 2004. [Online]. Available: http://www.jstor.org/stable/3318881

[40] A. Y. Ng and M. I. Jordan, "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes," in *Neural Information Processing Systems*, 2001, pp. 841–848.

[41] K. P. Bennett and E. J. Bredensteiner, "Duality and geometry in svm classifiers," in *In Proc. 17th International Conf. on Machine Learning*. Morgan Kaufmann, 2000, pp. 57–64.

[42] J. C. Waters, "Accuracy and precision in quantitative fluorescence microscopy," *The Journal of Cell Biology*, vol. 185, no. 7, pp. 1135–1148, 2009. [Online]. Available: http://jcb.rupress.org/content/185/7/1135.abstract