

Spring 2020

## From Cellular to Holistic: Development of Algorithms to Study Human Health and Diseases

Casey Anne Cole

Follow this and additional works at: <https://scholarcommons.sc.edu/etd>



Part of the [Computer Sciences Commons](#)

---

### Recommended Citation

Cole, C. A. (2020). *From Cellular to Holistic: Development of Algorithms to Study Human Health and Diseases*. (Doctoral dissertation). Retrieved from <https://scholarcommons.sc.edu/etd/5804>

This Open Access Dissertation is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact [dillarda@mailbox.sc.edu](mailto:dillarda@mailbox.sc.edu).

From Cellular to Holistic: Development of Algorithms to Study Human Health and  
Diseases

by

Casey Anne Cole

Bachelor of Science  
University of South Carolina, 2015

Master of Science  
University of South Carolina, 2018

---

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Computer Science

College of Engineering and Computing

University of South Carolina

2020

Accepted by:

Homayoun Valafar, Major Professor

Duncan Buell, Committee Member

Marco Valtorta, Committee Member

Stephen Fenner, Committee Member

Joseph Flora, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

© Copyright by Casey Anne Cole, 2020  
All Rights Reserved.

## Acknowledgements

I would like to acknowledge my family, my advisor as well as the excellent professors at the University of South Carolina for their unwavering support throughout my education. Thank you.

## Abstract

The development of theoretical computational methods and their application has become widespread in the world today. In this dissertation, I present my work in the creation of models to detect and describe complex biological and health related problems. The first major part of my work centers around the creation and enhancement of methods to calculate protein structure and dynamics. To this end, substantial enhancement has been made to the software package REDCRAFT to better facilitate its usage in protein structure calculation. The enhancements have led to an overall increase in its ability to characterize proteins under difficult conditions such as high noise and low data density. Secondly, a database that allows for easy and comprehensive mining of protein structures has been created and deployed. We show preliminary results for its application to protein structure calculation. This database, among other applications, can be used to create input sets for computational models for prediction of protein structure. Lastly, I present my work on the creation of a theoretical model to describe discrete state protein dynamics. The results of this work can be used to describe many real-world dynamic systems. The second major part of my work centers around the application of machine learning techniques to create a system for the automated detection of smoking using accelerometer data from smartwatches. The first aspect of this work that will be presented is binary detection of smoking puffs. This model was then expanded to perform full cigarette session detection. Next, the model was reformulated to perform quantification of smoking (such as puff duration and the time between puffs). Lastly, a rotational matrix was derived to resolve ambiguities of smartwatches due to position of the watch on the wrist.

## Table of Contents

Acknowledgments .....	iii
Abstract.....	iv
List of Tables .....	viii
List of Figures.....	x
Introduction.....	1
Chapter 1: Introduction to Protein Structure and Dynamics .....	8
1.1 Proteins .....	8
1.2 Existing Methods of Structure Calculation/Characterization of Dynamics .....	9
1.3 Residual Dipolar Couplings.....	12
1.4 Previous Work in Utilizing RDCs in Calculation of Protein Structure and Dynamics.....	13
1.5 Research Objectives .....	16
Chapter 2: Contributions to the Study of Protein Structure and Dynamics .....	22
2.1 Increased Usability, Algorithmic Improvements and Incorporation of Data Mining for Structure Calculation of Proteins with REDCRAFT Software Package .	23
2.2 PDBMine: A Reformulation of the Protein Data Bank to Facilitate Structural Data Mining.....	43
2.3 Structure Calculation and Reconstruction of Discrete State Dynamics from Residual Dipolar Couplings.....	59

Chapter 3: Summarization of Major Contributions for the Creation and Improvements of Methods to Calculate Protein Structure and Dynamics .....	101
3.1 Improving the Usability of REDCRAFT.....	101
3.2 Data Driven Dihedral Angle Restraints.....	102
3.3 Creation of RDC-based Model of Protein Dynamics .....	103
3.4 Suggestions for Future Work.....	104
Chapter 4: Introduction to Smoking Detection.....	105
4.1 Smoking Behavior and How it is Traditionally Studied.....	105
4.2 Previous Work .....	106
4.3 Research Objectives .....	107
Chapter 5: Contributions to the Study of Automated Smoking Detection.....	112
5.1 Recognition of Smoking Gesture Using Smart Watch Technology .....	113
5.2 Detecting Smoking Events Using Accelerometer Data Collected Via Smartwatch Technology: A Feasibility Study .....	131
5.3 Clinical Quantification of Smoking Topography Using Smartwatch Technology .....	154
5.4 Resolving Ambiguities In Accelerometer Data Due To Location Of Sensor On Wrist In Application to Detection of Smoking Gesture .....	174
Chapter 6: Summarization of Major Results for the Automated Detection of Smoking .....	186
6.1 Puff Level Binary Detection of Smoking .....	186
6.2 Session Level Detection of Smoking.....	187
6.3 Automated Characterization of Smoking Topography.....	188

6.4 Resolving Ambiguities in Accelerometer Data Based on the Position of Smartwatch on Wrist .....	189
6.5 Suggested Future Work .....	189
Chapter 7: Conclusion.....	191
References.....	192
Appendix A: Discussion of Decimation .....	213
Appendix B: List of Publications.....	214
Appendix C: Permission for Reprint.....	219



## List of Tables

Table 2.1. Results for each of the datasets is shown.....	41
Table 2.2. Different modes of dynamics.....	66
Table 2.3. Order parameters used for the complex 2-state model of dynamics.....	76
Table 2.4. Order parameters used for the simulated 2-state arc motion and the simulated DGCR8 dynamics. ....	76
Table 2.5. Order parameters used for the complex 3-state model of dynamics.....	76
Table 2.6. Results of 2-State 60°, 30° and 15° arc motion experiments.....	89
Table 2.7. Results for 2-state complex dynamics experiments.....	90
Table 2.8. Results for 3-state dynamics experiments. ....	92
Table 2.9. Results in recovery of DGCR8 discrete state dynamics. ....	93
Table 2.10. Results for modeling of a 3-state dynamic as a 2-state.....	95
Table 2.11. Results for simulating 2-state dynamics in our 3-state dynamic equation. ...	96
Table 5.1. Summary of data utilized in analyses. ....	118
Table 5.2. Values for the TPR were calculated by iteratively excluding sessions from the four categories producing false negatives.....	150
Table 5.3. Values for the FPR were calculated by iteratively excluding sessions from the two categories producing false positives. ....	151
Table 5.4. A short description and shorthand designation are given for each position on the wrist. "Right side up" denotes that the watch is positioned such that text on the screen is not upside down. (L/R) denotes either the left (L) or right (R) wrist.....	177
Table 5.5. The values for $\alpha$ , $\beta$ and $\gamma$ are given for each of the configurations.....	179

Table 5.6. Accuracies in detection of individual smoking gestures inside of continuous sessions before and after transformation are reported. .... 184

## List of Figures

Figure 1.1 An example backbone of an amino acid is shown (red).....	9
Figure 1.2. Sensitivity of NOEs versus RDC restraints.....	11
Figure 2.1. An example of equivalent a) NEF RDC file and b) legacy REDCRAFT RDC file. ....	30
Figure 2.2. The main REDCRAFT GUI implemented in Qt5.....	36
Figure 2.3. Three examples of dialogues that can be triggered by REDCRAFT at various stages of its analysis.....	36
Figure 2.4. Computed structure of 1A1Z with 4Hz of experimental error a) produced by the legacy version, and b) by the improved decimation procedure. ....	38
Figure 2.5. A comparison of the structural similarity between the X-ray structure of 1A1Z and the computed structure of the entire structure by REDCRAFT using new RDC vectors and the NEF format. ....	38
Figure 2.6. Alignment for the structure with dihedral restraints (magenta), without (blue) and x-ray structure of ubiquitin (green) for the first set of RDCs ([C'-H, N-H]x2).....	40
Figure 2.7. Alignment for the structure with dihedral restraints (magenta), without (blue) and x-ray structure of ubiquitin (green) for the second set of RDCs (N-Hx2).....	40
Figure 2.8. Alignment for the structure with dihedral restraints (magenta), without (blue) and x-ray structure of ubiquitin (green) for the second set of RDCs (N-Hx1).....	41
Figure 2.9. Database schema for PDBMine.....	47
Figure 2.10. The abundance of each amino acid found in Uniprot (blue) and PDBMine (red).....	51
Figure 2.11. The general shape of the distribution of 2-mers in the database. ....	51
Figure 2.12. R-Space (PDBMine on left and accepted on right) for a) GLY and b) PRO. ....	52

Figure 2.13. R-Spaces for a) GLY of GLY-PRO, b) PRO of GLY-PRO, c) first PRO of PRO-PRO and d) second PRO of PRO-PRO. ....	54
Figure 2.14. R-Space for the triplet GLY-PRO-PRO a) for GLY, b) for first PRO, c) for second PRO.....	55
Figure 2.15. Examples of the differences for residue 14 of ubiquitin at lengths of k= 3, 6 and 7.....	56
Figure 2.16. Resulting structures for k=3,6,7 (red, green and purple) aligned to the x-ray structure (blue).....	56
Figure 2.17. Example of a typical dynamic-profile for the protein 1A1Z in the absence of internal dynamics with simulated $\pm 1$ Hz of uniformly distributed noise.....	70
Figure 2.18. A diagram of illustrating our strategy for simultaneous characterization of structure and dynamics using REDCRAFT.....	72
Figure 2.19. 2-state arc motion of the protein 1A1Z by $60^\circ$ perturbation of the $\phi_{71}$ dihedral at residue 71.....	77
Figure 2.20. 2-state complex motion created by altering the dihedral angles of the protein 1A1Z at residue 58.....	78
Figure 2.21. 3-state complex model of dynamics with blue representing the static domain and the dynamic domain shown in red, green and orange correspond to the conformational states 1, 2 and 3 respectively. ....	79
Figure 2.22. Two conformations of DGCR8 generated from Molecular Dynamics Simulation. RBD1(the static domain) is shown in blue, the linker region in purple and the two states of RBD2 are shown in red and green.....	81
Figure 2.23. An example of the dynamic-profile for (a) 2-state model of dynamics and (b) 3-state model of dynamics .....	84
Figure 2.24. Dynamic profile of a sample (a) Rigid-body dynamics and (b) Uncorrelated dynamics. ....	86
Figure 2.25. Reconstructed states from the DGCR8 experiment. Opaque renderings denote the fragments of the protein reconstructed. (a) The first conformation of the target structure is shown in red and the corresponding conformation is shown in yellow. (b) The second conformation of the target protein is shown in green and the corresponding reconstructed conformation is shown in orange. ....	94

Figure 2.26. The resulting conformations from forced modeling of a 2-state dynamic as a 3-state are shown here. Fragments shown in red and green correspond to the two actual conformational states while yellow depicting the phantom irrelevant conformation with 1% relative occupancy. .... 97

Figure 5.1. Overlay of all single smoking gestures with the X dimension in blue, Y in red and Z in green. .... 116

Figure 5.2. Overlay of all patterns collected for the following non-smoking gestures: (a) drinking, (b) scratching one's nose, (c) yawning, (d) coughing, (e) brushing hair behind one's ear, and (f) rubbing one's stomach. .... 117

Figure 5.3. Example of continuous smoking session..... 118

Figure 5.4. Accuracy, specificity and selectivity of the neural networks during training. The bars are individually labeled based on their respective training sets. .... 122

Figure 5.5. Accuracies, specificity and sensitivity in the individual gesture detection trials..... 123

Figure 5.6. Total number of false positives created by each non-smoking gesture. Each segment is labeled based on the dimension of the accelerometer data being used. .... 124

Figure 5.7. Example of continuous smoking session superimposed with the output of the neural network trained on the X dimension. .... 126

Figure 5.8. Averaged accuracies, specificities and sensitivities across the five continuous smoking gesture detection experiments with error bars representing the respective min and max of each value..... 126

Figure 5.9. Accuracies for five continuous non-smoking session trials. .... 127

Figure 5.10. Output of the neural network for the X dimension superimposed to the original smoking session..... 128

Figure 5.11. Results for the Pebble smart watch. .... 129

Figure 5.12. An illustration of accelerometer axes on a typical smartwatch is shown... 135

Figure 5.13. An example of a smoking session is shown. Each dimension of the accelerometer data is shown in blue, red and yellow. An ideal output of the ANN is shown in purple where each bump denotes a smoking gesture. .... 138

Figure 5.14. Examples of the following non-smoking sessions: (a) drinking, (b) eating, (c) walking, and (d) typing on a computer.....	140
Figure 5.15 Model of a smoking session: a) Puff duration > 0.75 seconds, b) Maximum rest time between puffs < 4 minutes and minimum rest time > 2.5 seconds, c) Minimum number of puffs in a session = 3 puffs, d) Session duration < 8 minutes. ....	142
Figure 5.16. A noisy non-smoking session is shown before the smoothing filter with the output of the detection mechanism shown in purple. ....	145
Figure 5.17. A noisy non-smoking session is shown after the smoothing filter with the output of the detection mechanism shown in purple. ....	145
Figure 5.18. This session was not reported by the participant but is an unmistakable smoking session with 13 clear puffs. ....	147
Figure 5.19. This session was reported as a smoking session, but no clear smoking gestures can be identified.....	147
Figure 5.20. Outline of the protocol used for collection of in laboratory data. ....	161
Figure 5.21. A sample of ASPIRE's recording session illustrated in the upper right corner. The main and larger figure depicts the portion of the image that corresponds to a smoking session (asterisk indicate the start of a puff).....	163
Figure 5.22. Comparison of individual puff durations collected via CReSS (blue) and the App (orange). ....	165
Figure 5.23. Comparison of individual IPIs collected via CReSS (blue) and the App (orange). ....	165
Figure 5.24. Comparison of the visual puff count versus the a.) ASPIRE puff count and b.) CReSS puff count. In both figures' participant P14 was an outlier and is colored red. ....	166
Figure 5.25. Comparison of the overall a.) CReSS reported median puff duration v. ASPIRE reported and b.) CReSS reported median IPI and ASPIRE reported. ....	167
Figure 5.26. An illustration of smoking topography reported by CReSS, ASPIRE, and corrected CReSS. The puff durations and IPIs are illustrated in green and blue respectively. ....	168
Figure 5.27. Correlations of individual puff durations collected via the CReSS device and ASPIRE for participants a.) P15, b.) P17, c.) P19 and d.) P8. ....	169

Figure 5.28. Correlations of individual IPIs collected via the CReSS device and ASPIRE for participant a.) P15, b.) P19, c.) P7 and d.) P17. ....	169
Figure 5.29. Correlation of median puff volume and median puff duration.....	170
Figure 5.30. XYZ plots for configuration: (a) LP1, (b) LP2, (c) LP3, (d) LP4, (e) RP1, (f) RP2, (g) RP3, and (h) RP4.....	177
Figure 5.31. XYZ plots for rotated positions: (a) LP1, (b) LP2, (c) LP3, (d) LP4, (e) RP1, (f) RP2, (g) RP3, and (h) RP4.....	180
Figure 5.32. The number of correctly identified smoking gestures before transformation is denoted by the blue bars. The signals after transformation are represented by the red bars.....	182
Figure 5.33. Continuous session with watch in RP1 configuration (a) before and (b) after transformation is depicted. Output of the XYZ-ANN is shown in black. ....	184

## Introduction

From the cars that we drive to the watch that we wear on our wrists, the world today is dominated by computers. This new technological age relies heavily on the algorithms and hardware that run these systems. In addition, the application of computer science is quickly spreading to every field, whether it be biochemists using sequence alignment algorithms or psychologists needing to analyze complex datasets. In response to this new demand, the field of computer science has evolved. There are now three major components of computer science: theory, computational or algorithmic development and the application of computer science, specifically artificial intelligence (AI). Whereas my interest has spanned all three of these areas, the work that I present here focusses more on the last two components: algorithmic design or taking theory and implementing it to solve real-world problems, and the application of artificial intelligence specifically in the realm of biological and health related problems.

Over three decades of research conducted internationally, have concluded that complete models of human health and diseases must comprise complex interactions of biological, behavioral, and environmental factors. Biological interactions can be measured on three main levels: molecular, systems and holistic. The study of molecules presents a unique challenge for the development of algorithmic solutions as well as applications of AI. A specific type of molecule, macromolecules, are responsible for many of the body's vital functions. Many technical advances have pushed the study of biological systems



forward[1, 2], however there is still room for additional improvements. For instance, the study of protein structure and dynamics is inherently limited by the traditional experimental methods of characterizing protein structure and dynamics. Specifically, proteins that undergo dynamics (which may be essential to carry out their function) are not able to be adequately studied by the “gold standard” of structure determination, x-ray crystallography. To overcome this limitation, other experimental methods have been developed. One such method, Nuclear Magnetic Resonance spectroscopy (or NMR) has shown great promise in its ability to collect data that does not only allow for characterization of high-resolution protein structures, but also the characterization of dynamics. However, due to the complexities of the data collected, NMR has not yet realized its full potential in the study of proteins. In these instances where experimental methods are limited, computational methods can be developed to help fill the knowledge gap. A more complete understanding of protein structure and dynamics could have profound effects in the characterization of diseases and development of treatments. Therefore, the first focus of this work is to address existing gaps in the study of macromolecules, especially proteins.

While there have been substantial technological advances in studying the biological and environmental basis of diseases, there have been relatively minor advances in technologies for characterizing human behavior, *especially in context*. Given the relationship between environmental factors and human behavior, it can be argued that better understanding of the contextual and social aspects of behavior will lead to more effective and personalized interventions to promote healthier behaviors, such as smoking cessation, weight loss, and exercise. For example, cigarette addiction is a chronic, relapsing

brain disorder and remains the leading preventable cause of death and disability in the US and costs nearly \$200 billion each year. Although ~20% of adults in the USA currently smoke, the majority want to quit, and among those that make a quit attempt, the majority relapse. Human behavioral and neuroimaging studies in the laboratory have been the predominant method for studying smoking behavior, i.e., smoking in the lab or prediction of self-reported relapse in real word contexts. Methods for characterizing precipitants to and engagement in smoking behavior outside the laboratory remains under-developed.

Traditional approaches to the study of human behavior are limited to self-reporting or laboratory observations. While both research approaches have provided substantial insight into behavior, they suffer from some inherent limitations. Self-reporting is limited by recall bias and social desirability, among other issues that reduce internal and external validity. While laboratory-based studies are usually designed to overcome such biases, they may lack “ecological validity,” or the findings may fail to generalize to real-world settings. Because of this restricted insight about behavior, interventions to promote healthy behavior often fail. Therefore, to address the lacuna of data on factors influencing behavior, enabling participatory technologies need to be developed to allow non-intrusive and continuous observation of behavior in context and in natural settings. Mobile devices, which are widely used and include a rich array of sensors, can be a powerful platform for the development of methods for studying and influencing behavior.

Reports indicate more than 80% of the adults residing in urban areas in the US possess smartphones (>60% in rural areas). In the meantime, the sales and use of smartwatches have monotonically increased in the US and the world, indicating some certainty about their continued growth and use nationally and internationally. Relinquishing the need for

any wired infrastructure is especially helpful in less developed countries and rural areas. In addition, the use of smart devices is not confined to any particular socioeconomic class. Therefore, the use of smart and wearable devices for sensing and delivery of intervention approaches has the potential for international deployment without socioeconomic, political, or geographical barriers.

While these mobile technologies have pervaded and revolutionized much of our social and private lives, their utilization in healthcare remains sparse. Although the use of smart devices and web-based study of behavior have gained traction in recent years, additional human studies need to be conducted to establish the efficacy and usability of such devices in the study of behavior in natural settings. Effective use of these devices in health-promotion requires solving the following crucial problem: how to continuously monitor and characterize behavior in a way that will be accurate, complete, unobtrusive, relevant, personalized, and initiate the optimal intervention method. To this end, enabling technologies can be developed to allow real-time remote sensing of data collected from participants and interpreted by an artificial intelligence.

Solving these challenges call for usage of all three main components of computer science. First, a theoretical understanding and formulation of the problem must be established. Next, design of algorithms that models this formulation to solve specific problems. Finally, the incorporation of AI techniques to enhance the results achieved from the application of the model. The first objective of this work was to develop methods of analyzing data from NMR spectroscopy to predict both structure and dynamics of proteins. The second objective was to eliminate the barriers standing in the way of employing modern technology to monitor and study human behavior in situ, passively, accurately, and

intelligently. Developments will also establish the foundation to initiate automated and just-in-time interventions as needed.

The work presenting in this dissertation is structured into two distinct, but related parts. Both parts utilize various aspects of computer science to solve a variety of biological and health-related problems. Part 1 presents and details my contributions to the problem of solving protein structure and dynamics. My contributions can be broadly summarized as follows:

- Substantial improvements were made to the existing software package REDCRAFT including enhancement to its decimation routine, creation of a graphical user interface, refactoring of the base code to meet industry standards, and incorporation of dihedral angle restraints mined from a newly created database. All of these improvements have led to an improved user experience as well as an increase its ability to accurately characterize protein structure in a reasonable time scale.
- The protein databank (PDB) was dissected and used to create the PDBMine database. This database allows for easy and flexible mining of the data contained within over 400,000 protein structures. The results from various queries of this database have been shown to greatly enhance protein structure prediction methods.
- A method for characterizing atomic-level description of discrete state dynamics of proteins was created and tested using a variety of models. This model showed success on 2-state and 3-state models with motions as small as 15 degrees and rates of occupancy as low as 10 percent.

Part 2 of this document presents my work in the creation of automated systems to detect

smoking using smartwatch technology and AI. My contributions to this end can be broadly summarized as follows:

- An artificial neural network was used to create a predictive model to perform binary classification of a smoking puff using accelerometer data from a smartwatch. This model was tested on a variety of different non-smoking gestures and was able to detect smoking with over 80% accuracy.
- A theoretical model of a smoking session was established and, in combination with an artificial neural network, was used to study a cohort of 10 smokers. The system achieved over 90% accuracy in the detection of smoking sessions while maintaining a low false positive rate (<2%). The theoretical model of smoking serves as the first of its kind. It and the ANN were incorporated into a smartwatch application, ASPIRE, aimed at promoting cessation.
- The efficacy of using a smartwatch-based application instead of the expensive CReSS device to measure smoking topography such as puff duration and inner puff interval was investigated. The smartwatch-based application included an ANN that detects the following mini-gestures: hand-to-mouth, hand-on-mouth, hand-off-mouth and non-smoking. Based on the mini gestures detected, smoking topographies are measured. The various measured topographies exhibited  $R^2$  values ranging from 0.7-0.98 when compared to measurements taken from the CReSS device.
- The issue of rotational ambiguities arising from the position of the smartwatch on the wrist was resolved by deriving a set of Euler rotations. This rotation matrix was tested with six of the most probably positions of a watch on the wrist.

Each part of this document starts with an introductory chapter, followed by a chapter containing a short description of each research objective and all pertinent publications and is concluded with a summarization of the major results from each part.

## Chapter 1: Introduction to Protein Structure and Dynamics

The following subsections describe a brief overview of concepts vital to the understanding of the work presented. The following concepts will be covered: protein structure, existing methods for characterization of proteins, residual dipolar couplings (RDCs) as well as an overview of existing methods that utilize RDCs.

### 1.1 Proteins

Proteins are a class of polypeptides that perform various functions in cells including structural support, enzymatic activities, cell signaling, and many more. Production and retention of proteins is crucial for life in all organisms. The three-dimensional structure of a protein, or its “fold”, is one factor in determining how a protein functions. Structural variations can lead to disruption of a protein’s ability to function. These disruptions have been shown to cause a variety of diseases. In addition to the overall structure of a protein, some proteins undergo dynamics (hTS, DHFR, Hemoglobin, Myoglobin, MBP, SRD10 to name a few). In many cases, this motion is crucial for the protein to adequately function while in others, the internal dynamics of the protein marks the on-set of a disease.

Three dihedral angles, phi ( $\Phi$ ), psi ( $\Psi$ ), and omega ( $\Omega$ ), define the structural variability of the protein backbone (shown in red in Figure 1.1) at each amino acid.

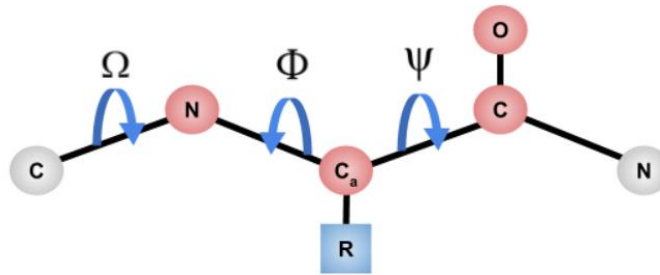


Figure 1.1 An example backbone of an amino acid is shown (red).

The collective effect of these torsion angles defines the overall structure of a protein. Due to the biophysical properties of the peptide bond, the  $\Omega$  torsion angle is generally fixed at  $180^\circ$  (or occasionally  $0^\circ$ ). Therefore, the effective degrees of freedom at any given amino acid are  $\phi$  and  $\Psi$ .

## 1.2 Existing Methods of Structure Calculation/Characterization of Dynamics

There are three main methods of protein structure determination and characterization of dynamics: x-ray crystallography, NMR spectroscopy, and computational methods. The first two methods are performed in a “wet” lab setting. Computational methods, on the other hand, are conducted in-silico and comprise of many different techniques including ab initio, threading and model-based. As stated previously, some proteins undergo dynamics that is critical in their function. Therefore, study of a particular protein must always include a study of any potential internal dynamics within the structure. Several studies have shown that ignoring the potential dynamics of a protein could lead to prediction of an erroneous structure[3-5].

X-ray crystallography is the “gold standard” of protein structure calculation. In this method the protein is desiccated in a crystalline setting and characterized. Whereas this method has led to many high-quality structures, it has some inherent limitations. First, the



decimation process is often very time consuming and therefore expensive. Some types of proteins, such as proteins embedded in structures such as the cellular membrane, are not able to be characterized in this way. In addition, proteins with dynamics are unable to be characterized due to the fact that in order for successful characterization the structure must be completely immobilized. Despite this limitation, some dynamical proteins have been studied by x-ray crystallography by artificially “locking” the protein into each configuration. However, by artificially locking the protein into various states has several practical limitations.

An alternative experimental approach is NMR spectroscopy in which the protein is suspended in a more native, liquid environment. The traditional data that is acquired from NMR are called NOEs, an acronym for Nuclear Overhauser Effect. These are distance restraints between pairs of atoms. These distance restraints are then used with a variety of programs to calculate protein structure. This method also results in high-quality structures, very similar to x-ray structures. However, NOE data on its own is not sufficient to determine the motion of proteins, but another restraint, residual dipolar couplings, collected via NMR holds the promise of being able to provide both a high-quality description of protein structure as well as any dynamics it is undergoing. Figure 1.2 shows the sensitivity of RDCs over NOEs. In this figure, the y-axis represents fitness to experimental data and the x-axis represents structural similarity to a target structure. Note that structural similarity of under 5 angstroms is generally considered to be significant. All dots represent a distinct variation of a protein structure. These structures were generated from a seed structure by randomly altering its structure. The blue dots represent the structural similarity of the mutant structures to the seed structures as a function of RDC fitness (or how well the data

fits the structure). The red dots represent the structural similarity of the same mutant structures as a function of the NOE fitness. Notice that the RDC fitness converges much more quickly than that of the NOE fitness. This indicates that a structure over 7 angstroms away from the target structure has the same NOE fitness as one that is within 1 angstrom of the true structure. This could cause NOE-based calculations to mistakenly choose the wrong structure due to its good fitness to the data. As shown in the figure, this differs with RDCs. A structure that is over 7 angstroms away from the native structure has very poor fitness score and therefore would not be chosen as the true structure in a blind study.

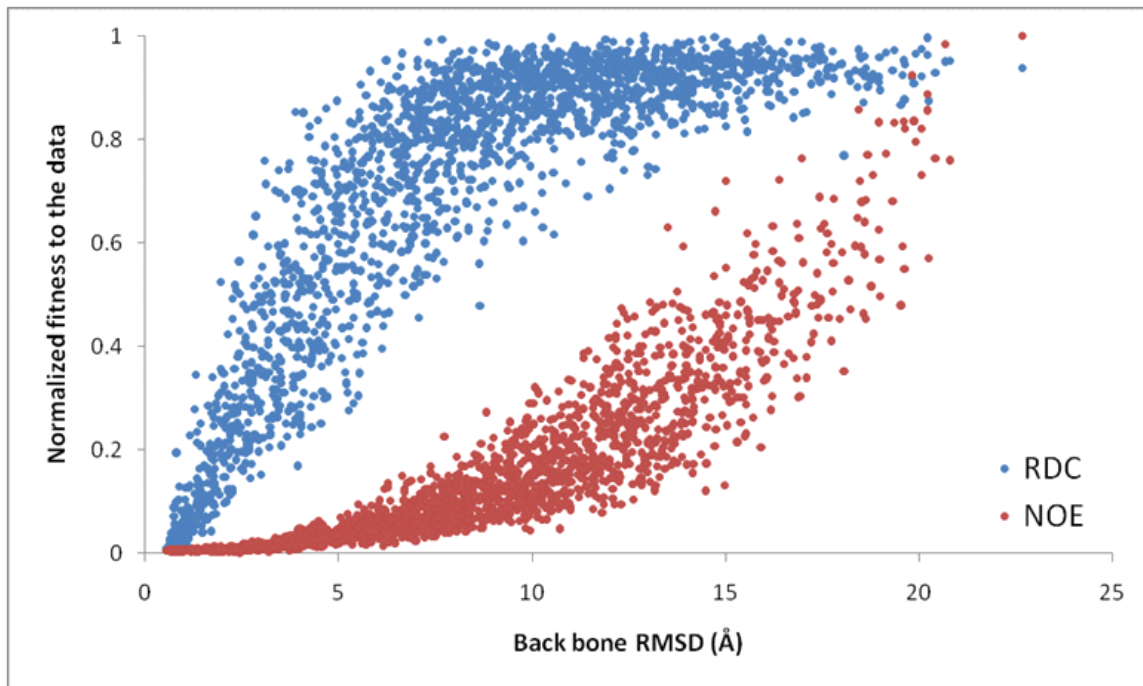


Figure 1.2. Sensitivity of NOEs versus RDC restraints.

The final class of approaches to protein structure modelling and characterization of dynamics is computational methods. These approaches are appealing for many reasons mainly a substantial decrease in cost and time required for completion. There are three main types of methods: model-based, threading and ab initio. In model-based structure

calculation, a pre-existing protein structure is used as a starting point. The structure is then refined based on energy calculations or fitness to experimental data. This method makes the assumption that a similar structure is available that can be used as a starting point. For novel structures, this is rarely the case. Threading methods improve on this process by partitioning the protein into smaller sequences and then finding model fragments from existing structures. When enough of these structures are found then they are combined, or “threaded” together to create the full structure. This method does not assume a full model structure is available but does assume that the individual fragments exist. This is more likely than finding the full structure but is still difficult in some instances. Ab initio methods are the most computationally demanding of the three. They attempt to characterize the structure of a protein by using only the sequence of the protein and minimal additional information such as a description of physical forces and a sparse set of experimental data. No additional assumptions are made.

### 1.3 Residual Dipolar Couplings

RDCs arise from the interaction of two magnetically active nuclei in the presence of the external magnetic field of an NMR instrument. This interaction is normally reduced to zero due to the isotropic tumbling of molecules in their aqueous environment. However, the introduction of partial order to the molecular alignment by minutely limiting their isotropic tumbling will reintroduce the finite RDC interactions. The resulting RDCs are measured relatively easily and represent an abundant source of precise and informative data which includes the relative orientation of different internuclear bonds within the alignment frame. An RDC can be calculated for a given set of atoms using the Equation (1.3.1) where  $D_{\max}$  and  $r_{ij}$  are constants calculated using the properties of the atoms

involved,  $v_{ij}$  is the calculated vector between atom  $i$  and  $j$ , and  $S$  is the order tensor given in Equation (1.3.2).

$$D_{ij} = \left( \frac{D_{max}}{r_{ij}^3} \right) v_{ij} \cdot S \cdot v_{ij}^T \quad (1.3.1)$$

$$S = \begin{bmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{xy} & S_{yy} & S_{yz} \\ S_{xz} & S_{yz} & S_{zz} \end{bmatrix}, v_{ij} = \begin{pmatrix} \cos(\theta_x) \\ \cos(\theta_y) \\ \cos(\theta_z) \end{pmatrix} \quad (1.3.2)$$

RDCs have been shown to be potentially instrumental in structural characterization of aqueous proteins[6] and other challenging proteins, while enabling simultaneous study of structure and dynamics of proteins[7, 8]. Because RDCs can be used to characterize the structure and dynamics of challenging proteins, it presents a viable, cost-effective method with the benefit of producing rapid, comprehensive and automated results. However, despite their potential, only a handful of protein structures submitted to the Protein DataBank[9] (PDB) have been determined exclusively by RDC data. Of this scarce number of proteins, a significantly high number of them have utilized far more RDCs than necessary[7, 10] therefore mitigating the advantages of using RDCs. Other approaches utilize less RDCs by resorting to use of other information such as NOE and hydrogen bond restraints[11] for successful structure determination. The primary contributing factor for the failure to realize the full potential of RDCs is the lack of analysis strategies capable of extracting the pertinent information from RDCs.

#### 1.4 Previous Work in Utilizing RDCs in Calculation of Protein Structure and Dynamics

Due to their high potential to solve both structure and accurately characterize dynamics, the need for software that can analyze RDC data and extract useful information

is paramount. Computational approaches such as CHARMM[12], AMBER[13], GROMACS[14], or NAMD[15] provide simulations of molecular dynamics (MD). These platforms incorporate nearly all of the understood biophysical forces at the atomic level and have incorporated some rudimentary form of RDC restraints. However, these platforms are generally incomplete and therefore consistently report that structure calculation from RDCs alone is improbable. A few methods exist that allow for structure calculation from RDCs alone, or at least that rely more heavily on RDC data. These methods have shown minor successes in the calculation of small protein structures (< 60 residues). In addition, some of the MD simulation software that have incorporated RDCs have been able to utilize them for characterization of dynamics under some favorable conditions. In order to perform this characterization, proteins are subjected to long simulations. The results of these simulations are a collection of structures called an ensemble. These structures are then clustered, and a representative structure is taken from each cluster as a conformational state. Whereas these simulations can yield plausible results, they have been demonstrated to be successful under extremely circumstantial conditions. Furthermore, these approaches are time-consuming and are not applicable to all proteins, for example large proteins, due to the complexity of their structures. One potential alternative to characterization of dynamics is to develop an RDC based model of dynamics in which the relative order of two domains of a protein is compared and then related to one another as a dynamic ensemble. Due to the inherently precise orientation information that exists in RDC data, it is hypothesized that such an endeavor will be possible.

The software package REDCRAFT[16, 17] has been previously developed in the Valafar Group and employs RDCs to calculate protein structures and sets itself apart from

other existing software packages by deploying a new and more efficient search mechanism. As a result, REDCRAFT can achieve the same structure determination outcome as other methods with less data. REDCRAFT is designed for structure determination primarily from RDC restraints and it sets itself apart from other approaches in a number of ways. REDCRAFT is able to accomplish this improvement by analyzing a given molecule through rigid peptide planes which adhere to strict peptide geometry. Therefore, all variations within a molecule are described through backbone torsion angles ( $\phi$ ,  $\Psi$ ). In terms of computation, this significantly reduces the number of variables, translating to a reduction in the dimensionality of the search space and improvements in program execution time. It also contributes to the program's robustness, allowing for fragmented study of protein structures when segments of data are erroneous or missing. Applications of REDCRAFT have been demonstrated using aqueous[18, 19] and membrane[20] proteins with as little as two RDCs per residue. Yet another advantage afforded by the publicly available REDCRAFT software package is its development using a sound Object Oriented (OO) programming paradigm that is easily extendable. This lends itself well to encapsulation of physical and biophysical properties of proteins since construction of a Polypeptide object, from more fundamental Residue and Atom objects, directly reflects the natural process of protein polymerization. It also allows for better program source code readability and more efficient program development, while further contributing to improvements in execution time.

The first stage of REDCRAFT is of vital importance to the success of structure calculation. In this stage a list of possible dihedral angles is generated for the given amino acid sequence. Each possibility on this list is then ranked based on the structural fitness of

the local RDC data. From there the protein elongation process begins. Due to the robustness of the REDCRAFT algorithm, there are some instances where even if the true dihedral angles are not present in the list for a given residue, there is a possibility of recovery if there are enough RDCs present for residue. However, in large proteins only sparse sets of RDCs can be collected. Therefore, a more informed way of generating potential dihedral candidates will increase the capability of REDCRAFT to calculate larger, more complex proteins. Experimentally, dihedral angle restraints can be obtained using NMR. The program TALOS[21] can decipher this data and creating estimates of local dihedral angles. Although the angles produced by TALOS have been shown to be fairly accurate, it requires additional time and effort to collect more data. Data mining techniques can be potentially used as a substitute the angles generated by TALOS in structure calculation.

## 1.5 Research Objectives

The research objectives for this work are to 1) improve the usability of the REDCRAFT software package, 2) develop a data mining-based approach for the calculation of dihedral restraints and 3) create an analytical model of characterizing protein dynamics. The following subsections describe more details about each research objective.

### *1.5.1 - Improving the Usability of REDCRAFT*

Historically, REDCRAFT has utilized a text-based UI and has been limited to just the Linux/Unix operating systems. While the text-based UI is fairly complete, lack of a graphical UI or GUI has limited the widespread adoption of the method among non-computer scientists, many of whom are accustomed to interacting with a GUI in the Windows environment. Therefore, the first task in improving the usability of REDCRAFT was to design and implement a GUI. The platform for development was QT5 due to its

cross-platform availability. In addition, cmake was incorporated to make compilation across different systems as seamless as possible. The second improvement was to modify the format of the RDC data that is used by the software. The format style used by the previous version of REDCRAFT was too customized and did not allow for a more flexible set of RDCs to be used. In order to make the program more flexible in this sense, the format was changed to be in the NMR Exchange Format[22] (NEF). NEF is the internationally accepted data exchange format in the NMR community that will make creation of pipelines of NMR software seamless. In addition to making REDCRAFT more accessible to the NMR community, the NEF format also allows for a more flexible set of RDCs to be defined and used. The previous version of REDCRAFT only accepted six different RDC vectors that were hardcoded into the rigid input format. Alternatively, a NEF input file lets the user define the exact atoms that are involved in the RDC interaction allowing for different RDC vectors to be incorporated. Minor modifications to the REDCRAFT code were also made to ensure that the current model of a “residue” contains all the atoms for which the experimental RDCs are reported. To assess the modifications to the computational core, a detailed performance analysis was performed that measured the runtime, accuracy and space requirements before and after the changes. In addition, the new version of REDCRAFT was presented at various conferences in the NMR community in the form of poster presentations. The results of this work are reported in the section titled “2.1 Increased Usability, Algorithmic Improvements and Incorporation of Data Mining for Structure Calculation of Proteins with REDCRAFT Software Package” that is accepted, awaiting publication in the BMC Bioinformatics journal.



### *1.5.2 - Data Driven Dihedral Angle Restraints*

The second research objective was to improve the initial stage of the REDCRAFT algorithm by generating better predictions of starting points for the dihedral angles of each residue. To accomplish this task, existing databases of structures will be leveraged. The Protein DataBank (PDB) currently houses the atomic coordinates of around 150,000 protein structures. However, there was no easy way to extract information about the dihedral angles of these proteins. Therefore, the data contained in these structures were downloaded and transformed to create a minable database (PDBMine), in which data such as dihedral angles for a given fragment of residues can be easily extracted. Once the data transformation was complete, the Stage-I of REDCRAFT's algorithm was modified to be able to extract the information out of the database and create dihedral restraint files. Theoretically, a better starting point should yield a better result (one that is closer to the native structure). Structure calculation was performed on a collection of proteins varying in size and complexity both with, and without the inclusion of dihedral restraints. In all cases it was found that inclusion of the restraints mined from PDBMine quality of structures computed by REDCRAFT as well as allowed for calculation of structures using sparse datasets that are unheard of in the community of RDC-based researchers. The results of this work are reported in the section titled "2.2 PDBMine: A Reformulation of the Protein Data Bank to Facilitate Structural Data Mining" that was an accepted manuscript to the IEEE CSCI Conference in 2019. Additional results are also given in section 2.1 under the subheading: "Incorporation of Data-Driven Dihedral Restraints".

### *1.5.3 - Creation of RDC-based Model of Protein Dynamics*

The research objective was to accurately characterize discrete state dynamic models of proteins to the atomic level. This objective required three main steps. The first step was to accurately identify existence and mode of dynamics within a given protein. REDCRAFT was uniquely poised to accomplish this task due to its iterative elongation process. In theory, when a protein is being constructed with REDCRAFT the RDC fitness to the structure should grow at a steady rate and then plateau at a level consistent with the level of the experimental error in the data. However, when dynamics exists in a protein the RDC data that is collected is perturbed and there is an “averaging” of data between the multiple states, or conformations. In this case, when structure characterization is performed, assuming a single rigid structure, a spike in RDC fitness will be observed at the point in which the dynamics begins to alter the RDC data. The fitness of RDC data with respect to the residue number served as a profile of dynamics (dynamic-profile). This profile was investigated to ascertain its ability to characterize the different structural modes of dynamics, namely: temporal and structural. Along the structural mode of dynamics there are two categories of rigid-body and uncorrelated modes. Similar to previous definitions[23, 24], rigid-body dynamics is defined as dynamical regions that maintain a constant internal structure as a function of time, while the uncorrelated dynamics is defined as alteration of structure as a function of time. Therefore, it is meaningful to describe the structure of a dynamical region if it is engaged in a rigid-body dynamics, and not so for an uncorrelated mode of dynamics. The temporal dimension of dynamics can be defined by two categories of discrete-state and continuous-state dynamics. The distinction between the two is solely based on the temporal occupancy of conformational states that are visited

during the trajectory of the dynamics. In principle all four combined modes of dynamics should be possible with examples of all four combination of structural and temporal modes of dynamics having already been identified and presented in the literature. However, the combination of rigid-body, discrete-state dynamics represents the biologically most likely event and was therefore targeted in this work. It should also be noted that the remaining three modes (combinations of structural and temporal modes) can be approximated as a rigid-body and discrete-state dynamics in some favorable instances which makes this mode of dynamics all the more important to properly study.

After correctly identifying the regions of dynamics, the second step was to perform a fragmented characterization of the protein in which a portion of the static region was characterized separately from the dynamical region. In the final step, order tensors were calculated for each of the fragments and analytically compared to reconstruct the relative orientation of the  $n$  states, or conformations, of the protein.

The evaluation of this objective relied on simulated RDC data from distributions of closed and open states. Simulated data allowed for proper assessment of the algorithm against “ground truth” dynamic structures. They allowed precise quantification of the accuracy of the developments and helped to establish the limitations of the proposed approach as a function of experimental noise, missing data, and degeneracies of order tensors. Simulated data also allowed the investigation of the sensitivity of the approach with respect to all relevant parameters, including the quality of data, magnitude of motion, and level of occupancy. In addition to simulated models of dynamics, synthetic data was generated from experimentally confirmed models of dynamics. The method was then tested on these real systems and assessed for accuracy in comparison to the published

models. The results of this work are reported in the section titled “2.3 Structure Calculation and Reconstruction of Discrete State Dynamics from Residual Dipolar Couplings” that was an accepted manuscript to the Journal of Chemical Theory and Computation (JCTC) in 2016.

## Chapter 2: Contributions to the Study of Protein Structure and Dynamics<sup>1</sup>

---

<sup>1</sup> Publications in this chapter:

Section 2.1: Casey A. Cole, Caleb Parks, Julian Rachele and Homayoun Valafar. Accepted by BMC Bioinformatics. Reprinted here with permission of publisher.

Section 2.2: Casey A. Cole, Christopher Ott, Diego Valdes and Homayoun Valafar. 2019. Proceedings of 2019 IEEE International Conference on Computational Science & Computational Intelligence (IEEE CSCI). Reprinted here with permission from the publisher. See IEEE copyright: <https://www.ieee.org/publications/rights/copyright-policy.html> © 2019 IEEE

Section 2.3: Casey A. Cole, Rishi Mukhopadhyay, Hanin Omar, Mirko Hennig, and Homayoun Valafar. 2016. Published by J. Chem. Theory Comput. Reprinted here with permission of publisher.

## 2.1 Increased Usability, Algorithmic Improvements and Incorporation of Data

### Mining for Structure Calculation of Proteins with REDCRAFT Software Package

#### 2.1.1 Abstract

Traditional approaches to elucidation of protein structures by Nuclear Magnetic Resonance spectroscopy (NMR) rely on distance restraints also known as Nuclear Overhauser effects (NOEs). The use of NOEs as the primary source of structure determination by NMR spectroscopy is time consuming and expensive. Residual Dipolar Couplings (RDCs) have become an alternate approach for structure calculation by NMR spectroscopy. In previous works, the software package REDCRAFT has been presented as a means of harnessing the information containing in RDCs for structure calculation of proteins. However, to meet its full potential, several improvements to REDCRAFT must be made. In this work, we present improvements to REDCRAFT that include increased usability, better interoperability, and a more robust core algorithm. We have demonstrated the impact of the improved core algorithm in the successful folding of the protein 1A1Z with as high as  $\pm 4\text{Hz}$  of added error. The REDCRAFT computed structure from the highly corrupted data exhibited less than  $1.0\text{\AA}$  with respect to the X-ray structure. We have also demonstrated the interoperability of REDCRAFT in a few instances including with PDBMine to reduce the amount of required data in successful folding of proteins to unprecedented levels. Here we have demonstrated the successful folding of the protein 1D3Z (to within  $2.4\text{\AA}$  of the X-ray structure) using only N-H RDCs from one alignment medium. The additional GUI features of REDCRAFT combined with the NEF compliance have significantly increased the flexibility and usability of this software package. The

improvements of the core algorithm have substantially improved the robustness of REDCRAFT in utilizing less experimental data both in quality and quantity.

### *2.1.2 Background*

Faster and more cost-effective methods of characterizing protein structures are of paramount importance in the development of personalized medicine. While there have been substantial developments in reducing the cost, and increasing the speed of sequencing genomic data[25-28], there has been relatively little advances in improving the characterization of protein structures[29]. In addition to the existing disparity in genetic versus proteomic information, the vast majority of the characterized protein structures belong to a very specific and limited category of proteins. For instance, while it has been estimated that 30% of the human proteome consists of membrane proteins, this important class of proteins is represented by approximately 120 proteins in current databases[30, 31]. Such observed disparities are rooted in the lack of new approaches to structure calculation that overcomes the existing barriers in structural determination of proteins[32, 33].

In recent years, the use of Residual Dipolar Coupling (RDC) data acquired from Nuclear Magnetic Resonance (NMR) spectroscopy has become a potential avenue for a significant reduction in the cost of structure determination of proteins[31]. In addition, RDC data have been demonstrated to overcome some long-standing challenges in NMR spectroscopy such as structure determination of membrane proteins[20, 34-37], recognition of fold families[38] and the concurrent study of structure and dynamics of proteins[5, 7, 39-45]. Recent work[46-51] has demonstrated the challenges in structure calculation of proteins from RDC data alone, and some potential solutions have been introduced [47, 48, 51-54]. One such approach named REDCRAFT[20, 42, 46] has been demonstrated to be

successful in structure calculation of proteins from a reduced set of RDC data (and therefore reduced cost). While REDCRAFT has been very successful compared to other approaches, it exhibits some limitations that result in reduced usability and flexibility. In this work, we present usability and methodology improvements to REDCRAFT that aim to address these limitations. To increase the usability, we have incorporated a powerful Graphical User Interface (GUI), integrated it with molecular visualization software, and adopted the newly approved NMR Exchange Format[22] (NEF), to name a few. REDCRAFT's core methodology has been revised to allow calculation of protein structures under challenging conditions. More specifically, we improve the decimation routine as well as incorporate new dihedral restraints mined from the PDBMine[55] database. To evaluate the updates, we present and discuss structure calculation of proteins using novel sets of RDC data that REDCRAFT, under lower signal to noise conditions as well as with sparse sets of RDCs. The REDCRAFT package is purely developed in C++ according to valid software development principles and is freely available for download via Bitbucket repository (<https://bitbucket.org/hvalafar/redcraft/>).

#### *2.1.2.1 Residual Dipolar Couplings*

RDCs can be acquired via NMR spectroscopy and the theoretical basis of their interaction had been established and experimentally observed in 1963[56, 57]. RDC data has become a more prevalent source of data for structure determination of biological macromolecules in recent years due to the availability of alignment media[58] and substantial improvements in NMR instruments. Upon the reintroduction of order to an isotropically tumbling molecule, RDCs can be easily acquired. The alignment medium can impose restricted tumbling through steric, electrostatic, or magnetic interaction with the



protein. The RDC interaction between two magnetically active nuclei can be formulated as shown in Eq. (2.1.1).

$$D_{ij} = D_{max} \left\langle \frac{3\cos^2(\theta_{ij}(t)) - 1}{2} \right\rangle \quad (2.1.1)$$

$$D_{max} = \frac{-\mu_0 \gamma_i \gamma_j h}{(2\pi r)^3} \quad (2.1.2)$$

In this equation,  $D_{ij}$  denotes the residual dipolar coupling in units of hertz between nuclei  $i$  and  $j$ . The  $\theta_{ij}$  represents the time-dependent angle of the internuclear vector between nuclei  $i$  and  $j$  with respect to the external magnetic field of the NMR instrument, and the angle brackets signify time averaging. In Eq. (2.1.1),  $D_{max}$  represents a scalar multiplier dependent on the physical properties of the two interacting nuclei and is further described in Eq. (2.1.2). In this equation,  $\gamma_i$  and  $\gamma_j$  are nuclear gyromagnetic ratios of nuclei  $i$  and  $j$  respectively,  $r$  is the internuclear distance (assumed fixed for directly bonded atoms),  $h$  is the modified Planck's constant, and  $\mu_0$  is the permeability of free space. Additional description and alternate formulations of equations 2.1.1 and 2.1.2 can be found in the following work[46, 59, 60].

#### 2.1.2.2 REDCRAFT Structural Fitness Calculation

While generating a protein structure from a given set of residual dipolar couplings is nontrivial, it is straightforward to determine how well a given structure fits a set of RDCs. REDCRAFT's core approach utilizes this principle in order to produce a viable protein structure. Through algebraic manipulation of Eq. (2.1.1) RDC interaction can be represented as shown in Eq. (2.1.3),

$$D_{ij} = v_{ij} * S * v_{ij}^T \quad (2.2.3)$$

where  $S$  represents the Saupe order tensor matrix[9] and  $v_{ij}$  denotes the normalized interacting vector between the two interacting nuclei  $i$  and  $j$ . REDCRAFT takes advantage of this principle by quantifying the fitness of a protein to a given set of RDCs (in units of hertz) and calculating a root-mean-squared deviation as shown in Eq. (2.1.4). In this equation  $D_{ij}$  and  $D'_{ij}$  denote the computed and experimentally acquired RDCs respectively,  $N$ , represents the total number of RDCs for the entire protein, and  $M$  represents the total number of alignment media in which RDC data have been acquired. In this case, a smaller fitness value indicates a better structure.

$$Fitness = \sqrt{\frac{\sum_{j=1}^M \sum_{i=1}^N (D_{ij} - D'_{ij})^2}{M \times N}} \quad (2.1.4)$$

The REDCRAFT algorithm and its success in protein structure elucidation have been previously described and documented in detail[20, 46]. Here we present a brief overview. REDCRAFT calculates structures from RDCs using two separate stages. In the first stage (*Stage-I*), a list of all possible discretized torsion angles is created for each pair of adjoining peptide planes. This list is then filtered based on allowable regions within the Ramachandran space[56]. The list of torsion angles that remain is then ranked based on fitness to the RDC data. These lists of potential angle configurations are used to reduce the search space for the second stage.

*Stage-II* begins by constructing the first two peptide planes of the protein. Every possible combination of angles from *Stage-I* between peptide planes  $i$  and  $i+1$  are evaluated for fitness with respect to the collected data, and the best  $n$  candidate structures are selected, where  $n$  denotes the search depth. The list of dihedral angles corresponding to the top  $n$  structures is then combined with every possible set of dihedral angles connecting the next

peptide plane to the current fragment. Each of these candidate structures is evaluated for fitness and the best  $n$  are again selected and carried forward for additional rounds of elongation. All combination of dihedral angles worse than the best  $n$  are eliminated, thus removing an exponential number of candidate structures from the search space. This elongation process is repeated iteratively, incrementally adding peptide planes until the entire protein is constructed.

### *2.1.3 Implementation*

#### *2.1.3.1 Usability Updates to the REDCRAFT Software Package*

Several changes have been made to the REDCRAFT package to increase usability including reorganization, documentation, addition of a graphical user interface as well as adoption of NEF standards. These developments are outlined in the following subsections.

##### *2.1.3.1.1 Reorganization, Documentation and Addition of GUI*

The initial version of the REDCRAFT software package was only accessible through a Linux command line environment. Several changes have been incorporated to allow REDCRAFT to be mostly platform-agnostic, and it is now able to be compiled and executed on any Linux, BSD, or Unix system, including MacOS. Dependencies have also been updated such the latest version of the GNU C Compiler can be used for compilation. In addition, CMake[61] was integrated to all for dynamically generated makefiles that are suitable for an individual machine.

Regardless of the operating system, the command line environment could be cumbersome to use, especially for novice users. To create a more streamlined analysis pipeline, the project was reorganized to allow all REDCRAFT binaries and scripts to run from a single command instead of scattered individual pieces, thereby encapsulating the

project and facilitating simpler use. This is accomplished by only including a single binary, `redcraft` in the user's path that acts as command interpreter for the entire REDCRAFT project.

Additionally, a documentation system was put in place (<http://redcraft.readthedocs.io/>) that allows new documentation to be built and updated upon every update to REDCRAFT. This documentation details the steps necessary to compile the entire REDCRAFT suite, as well as dependencies. The documentation may be easily exported as HTML, DOCX, or PDF document formats for offline reference.

Finally, a modern Qt5 GUI system was developed to facilitate the usage of REDCRAFT even further. The GUI, written in C++ with Qt5, is fast and available uniformly across all platforms. The GUI contains tools to run *Stage-I* and *Stage-II*, reads config files, and allows for preliminary analysis of output files. Invocation of the GUI is performed by running either `redcraft gui` or `redcraft gui [path]` (to immediately launch the GUI in that directory).

#### *2.1.3.1.2 Adherence to NEF Standards*

The previous version of REDCRAFT utilized a rigid file format by allowing the analysis of only six specific RDC vectors (per residue) and their corresponding error values (example shown in Figure 2.1b). These six RDC classes represented the most prevalently collected vectors in the field of NMR at the time of REDCRAFT's creation. Since then, due to advances in instrumentation, introduction of new alignment media, and data acquisition techniques, a much wider range of RDCs can be collected to aid in structure calculation. To address issues such as this the NMR community introduced the NMR Exchange Format[22] (NEF). NEF is a standard for the representation of all NMR restraints

and accompanying data. NEF was created from a series of workshops and consultations with developers of NMR structure determination software developers to streamline the pipeline of structure determination programs. The NEF formulation of RDCs is much more flexible in its definitions (an example is shown in Figure 2.1a). NEF lists the name, residue number, and residue name of both atoms associated with each RDC along with the RDC value and uncertainty. To accommodate the robust possibilities of RDC values that NEF could contain, REDCRAFT's computational engine was expanded to handle any combination of the interacting nuclei along the backbone of a protein. The introduction of this standard has allowed the structure determination of proteins with data that was not possible before. To remain backward compatible, a conversion script is available that will convert the legacy format into the NEF format. This conversion script has also been integrated into the GUI.

<pre> PHE 4.05541322826833 PHE 2 PHE C 3 PHE N -3.03013 1 3 PHE N 3 PHE H 0.260408 1 2 PHE C 3 PHE H -2.49011 1 3 PHE CA 3 PHE HA -2.2099 1 3 PHE HA 3 PHE H -6.97504 1 2 PHE HA 3 PHE H 2.26648 1 LEU 6.18890572380312 LEU 3 LEU C 4 LEU N -1.07715 1 4 LEU N 4 LEU H 5.64809 1 3 LEU C 4 LEU H -2.13885 1 4 LEU CA 4 LEU HA -12.9958 1 4 LEU HA 4 LEU H 2.21004 1 3 LEU HA 4 LEU H 4.7 1 </pre>	<pre> PHE 4.05541322826833 PHE -3.03013 1 0.260408 1 -2.49011 1 -2.2099 1 -6.97504 1 2.26648 1 LEU 6.18890572380312 LEU -1.07715 1 5.64809 1 -2.13885 1 -12.9958 1 2.21004 1 4.7 1 </pre>
---	---

(a)

(b)

Figure 2.1. An example of equivalent a) NEF RDC file and b) legacy REDCRAFT RDC file.

### 2.1.3.2 Methodology Updates to the REDCRAFT Software Package

#### 2.1.3.2.1 Improvements of Decimation Methodology

REDCRAFT's core principle approach is to generate plausible structures in a combinatorial fashion and evaluate their fitness to the experimental data. To address the

intractability of combinatorial approaches, REDCRAFT has incorporated a static-decimation strategy (previously described in[46]) to reduce a large number of quasi-acceptable structures into a smaller and more manageable subset of structures by selecting representative structures. The static-decimation process utilizes user-specified parameters in order to balance the two competing objectives of examining a larger pool of structures versus the computational demands of a larger and more robust search for structures. Proper selection of these parameters is normally a simple process for typical data but becomes impossible for more noisy data. Consideration of structures with poor fitness to the data is unnecessary accommodation under high signal to noise ratio. However, under the conditions of low signal-to-noise ratio, the true structure is more likely to be subjected to early elimination based on poor fitness to the data.

The new version of REDCRAFT overcomes the limitation of the static-decimation process by introducing the more intelligent and adaptive dynamic-decimation process. In the dynamic-decimation process, the search and decimation parameters of REDCRAFT are automatically and dynamically adjusted at each stage of the analysis to reflect the quality, and therefore the computational demands of that stage. To accomplish this, a percentage threshold of tolerance ( $n$ ) is set instead of a static user-defined threshold. At each step in the elongation process of the algorithm, only structures with an RDC fitness score less than the current score of the fragment  $+n\%$  will be considered in the decimation pool. Using this new approach, two common and limiting impediments will be corrected. The first is in situations where there is low data density. In areas of low data density, the contribution from the static-decimation routine causes the solution space to grow exponentially, which is manifested in exponentially increasing computational resources (CPU and memory). For

example, during the first few steps of elongation there are typically a few RDCs, which result in underdetermined definition of the problem. In such instances a globally defined acceptance criterion would likely include nearly all the possible structural solutions, as all potential structures will have a low RDC fitness score. Dynamic decimation controls this intractable growth rate by only considering structures within  $n\%$  of the current score of the protein fragment. The second scenario appears in areas of highly noisy data. In these areas, contribution from decimation can drop to zero because of poor local structural fitness to the low-quality data. In this scenario, dynamic decimation assures controlled contribution from decimation.

Using the dynamic-decimation process we have investigated the low signal-to-noise instances of structure determinations that were not possible before. For this evaluation we have used the target protein 1A1Z, for which structure calculation has not been successfully completed using RDC data with low signal-to-noise ratios. In our experiments we have pushed the limits of the structure determination of this protein with as much as  $\pm 4$  hertz of added uniform noise.

#### *2.1.3.2.2 Incorporation of Data-Driven Dihedral Restraints*

The protein databank[62] (PDB) currently houses close to 150,000 protein structures. However vast this collection, the data storage format does not allow for easy mining of low-level information such as dihedral angles restraints. However, recently a minable version of the PDB has been created called PDBMine. Using PDBMine, a protein sequence and a rolling window size is inputted. The protein is then fragmented into k-mers using the rolling window. The dihedral angles are then extracted from these fragments and aggregated for a given amino acid and the most likely dihedral is predicted. The resulting

information can then be used to generate the candidate angle files created in *Stage-I* of REDCRAFT by varying the predicted angles  $\pm n$  degrees ( $n = 25$  in this work). In this work, datasets as low as one RDC per residue in only one alignment medium will be used to characterize ubiquitin. Ubiquitin (1D3Z) was chosen due to the availability of both high resolution RDC data as well as both x-ray and NMR structure to compare results. It has also been the subject of past RDC studies[40, 47, 63-66] that serve as comparisons for the results of this study. To date, there has been no successful attempt of structure characterization with this sparse of data.

### 2.1.3.3 Evaluation Protocol

Throughout the process of evaluating the new features of REDCRAFT we have utilized two target proteins 1A1Z and 1D3Z. These two proteins have been selected because they represent helical proteins, appropriate in size for study by NMR spectroscopy, and have been the subjects of previous studies by RDC data. Each of these proteins provide challenging cases. For example, 1A1Z is a difficult protein to characterize due to its helical nature[67] and structural anomalies that force it to sample atypical Ramachandran Space[68]. The protein 1D3Z also provide other unique challenges due to its helical nature and hypothesized internal dynamics. The helical proteins are generally more challenging to study by RDCs since the backbone N-H vectors are in nearly parallel configuration. The dynamical nature of 1D3Z protein will provide a challenging case of establishing its backbone dihedrals. Other additional challenging attributes of each protein that qualifies them for our studies are described in individual sections.

Our evaluation of REDCRAFT's improved decimation routine proceeded in three main steps. During the first step, the known structure of 1A1Z was used to generate



simulated RDC data using typical order tensors previously used in several studies[46, 69] and the software package REDCAT[70]. The RDC set simulated included the four of the previously available RDC vectors as well as two new vectors ( $[H^{\alpha}-C^{\alpha}, N-C^{\alpha}]$ ) that were previously unusable in REDCRAFT. Evaluation of a new methodology such as REDCRAFT based on simulated RDC data is of critical value. The use of simulated data allows for exact control over the quality of data, quantification of the performance as a function of signal-to-noise ratio, and proper assessment of time and space complexity of an algorithm as a function of data quality, to name a few.

In addition, to test the utility of incorporating data driven dihedral angles, the protein 1D3Z, the NMR structure of ubiquitin, was used. Due to the availability of experimental RDCs for 1D3Z, no additional synthetic RDCs were generated. Previous results for 1D3Z using REDCRAFT have shown that for high resolution structure calculation, at least two RDC vectors in two alignment media are required. To test the new dihedral restraints, we will attempt to decrease the total RDCs needed.

During the second phase of evaluation, the simulated RDC data are utilized by REDCRAFT to generate a protein structure. During this phase of the experiment, the REDCRAFT's RDC-fitness score was used to evaluate the success of REDCRAFT. If successful, the viable structures should exhibit an RDC-fitness to the data that is in the same order of the experimental error (related to the signal-to-noise).

Finally, during the third step, the computed structure is compared to the starting structure (the ground-truth) in order to ascertain the success of REDCRAFT. To evaluate structural similarity, the bb-rmsd (backbone root mean squared deviation) between resulting REDCRAFT structures and the target structure was calculated. The measure of

bb-rmsd is prevalently used to establish the structural similarity between two proteins. Values under 3.5Å can signify the success of REDCRAFT under noisy data conditions, while values under 2Å can be interpreted as strong evidence for structural similarity.

#### 2.1.4 Results/Discussion

##### 2.1.4.1 Integration of Graphical User Interface

The Graphical User Interface (GUI), written in Qt5, was integrated seamlessly into the REDCRAFT package utilizing CMake. Qt5 contains CMake bindings to link all the necessary Qt dependencies, therefore the end user will notice no difference between compiling the REDCRAFT engine and the GUI itself. The GUI can be launched directly from the command line so that it may immediately open the current working directory, or it may be launched from its binary. REDCRAFT and subsequently REDCRAFT GUI runs seamlessly on all flavors of Linux as well as macOS. Dependencies for this version of REDCRAFT are the GCC G++ Compiler, OpenMP (used for parallelization of processing), Qt5 with Charts (for GUI support), and Python 3 and Perl (for auxiliary script support). Instructions for installation of all dependencies can be found in the REDCRAFT documentation (<https://redcraft.readthedocs.io/>).

After executing the GUI, the user will be presented with the screen shown in Figure 2.2. The initial screen consists of four panels. The first panel (Panel A) displays a greeting message as well as some “quick tips” to aid the user in utilization. Panel B loads the run parameters for *Stage-I* and *Stage-II*. Tabs allow for easy navigation between the two stages. Panel C shows all files present in the user’s working directory, that is, the folder in which the REDCRAFT GUI was started in. This working directory can be changed via File->Open Directory at the top left of the GUI. In Panel D the output of each stage of execution

is printed. For instance, if the “Execute Stage 1” button is pressed then the results of Stage-I angle creation will be shown (see Figure 2.3a and 2.3b as examples). When in the “Stage 2” tab of Panel B, if the “Execute Stage 2” button is pressed then the results of Stage-II calculation will be shown in Panel D. When the “Advanced” tab is selected in the Stage 2 tab on Panel B, the panel expands to fill the entire column (as seen in Figure 2.3c) and additional parameters are shown. At any time during the execution of either stage, the process can be stopped by pressing the stage’s respective “Stop” button (shown in red on Panel B).

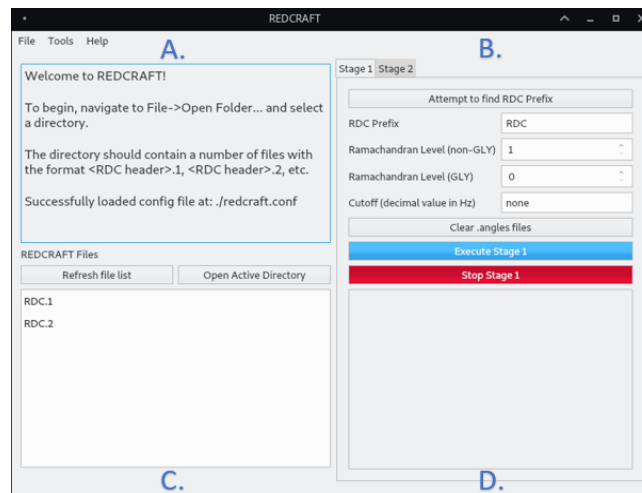


Figure 2.2. The main REDCRAFT GUI implemented in Qt5.

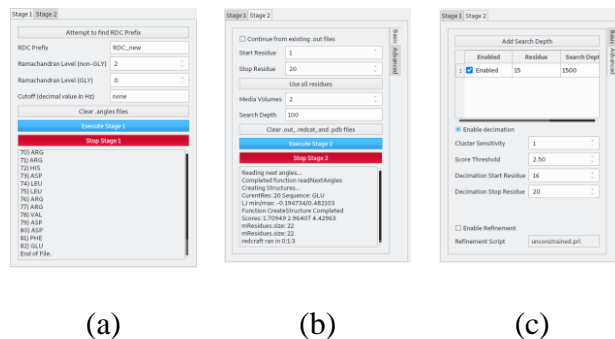


Figure 2.3. Three examples of dialogues that can be triggered by REDCRAFT at various stages of its analysis.

After executing the REDCRAFT analysis through its GUI, the resulting config file follows the standard INI format, but with comment support. The user is free to modify the configuration file directly, but the GUI will automatically eliminate any additional user comments in order to maintain backward compatibility.

#### *2.1.4.2 Results of Structure Calculation Using Improved Decimation Method*

The new version of decimation is universally faster than the previous version. Figure 2.4 shows the results of the first 20 residues of 1A1Z (using RDC data with  $\pm 4$  Hz of error) folded with the previous version of decimation compared to the same segment folded using the new decimation method using identical search parameters. The 20-residue (out of 83 total) segment of 1A1Z was selected due to the excessive space requirement of the previous version of decimation. The previous version required 4 hours of analysis time, at the end of which the final structure exhibited a bb-rmsd of 1.589Å to the reference structure (RDC fitness score of 2.21, results shown in Figure 2.4a). However, the extension of this fragment required memory in excess of the 16GB of the host computer and therefore did not complete the full analysis of the protein within a week. The new version of the decimation completed this exact segment on the same host computer in about 4 minutes and produced a structure with backbone bb-rmsd similarity of 0.946Å to the reference structure (RDC fitness score of 2.19, shown in Figure 2.4b). Of the greater importance is the success of the new version of REDCRAFT in providing a full structure of 1A1Z (illustrated in Figure 2.5 and discussed in the next section) that was never completed by the previous version of the software.

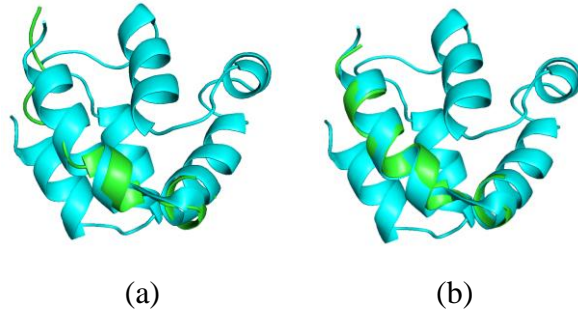


Figure 2.4. Computed structure of 1A1Z with 4Hz of experimental error a) produced by the legacy version, and b) by the improved decimation procedure.

#### 2.1.4.3 Results Reconstruction of Proteins Using NEF Format

The changes to the core REDCRAFT engine to accept NEF format enable it to perform the structure calculation of proteins based on a flexible set of RDC data. RDC pairs that were unavailable in the old version are now able to be used for reconstruction. For example, 1A1Z with  $[H^{\alpha}-C^{\alpha}, N-C^{\alpha}]$  RDC data in two alignment media with 0 Hz of simulated noise can now be folded with REDCRAFT. Using the new decimation approach, REDCRAFT produced the final structure of 1A1Z with a bb-rmsd of  $1.404\text{\AA}$  and an RDC fitness score of 0.835 when compared to X-ray structure of 1A1Z (Figure 2.5). This is a substantial achievement in the successful folding of a protein with flexibly defined RDCs.

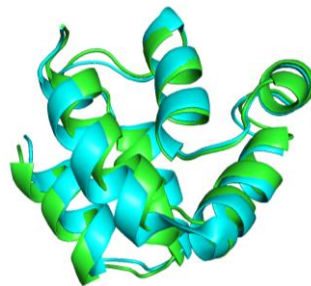


Figure 2.5. A comparison of the structural similarity between the X-ray structure of 1A1Z and the computed structure of the entire structure by REDCRAFT using new RDC vectors and the NEF format.

However, it should be noted that this modification causes a slight increase in runtime that can vary from 1-5% slower than the previous version. The time requirements were benchmarked by performing structure calculation of the same protein, using the same set of RDCs in both the previous and NEF-compatible version. Typically, the new version of REDCRAFT completes within a minute of the previous version for an analysis that takes approximately 45 minutes, and therefore the slower performance is considered negligible.

#### 2.1.4.4 Incorporation of Data-Driven Dihedral Restraints

The protein sequence for ubiquitin (76 residues) was submitted to PDBMine with a rolling window size of six. The resulting dihedral predictions for each amino acid was then used to create dihedral restraints by varying them +/- 25 degrees in steps of 5 degrees to be used in *Stage-I* of REDCRAFT. The structure was then calculated with a varying set of RDC data both with and without the PDBMine-based dihedral restraints. For each set of data, a figure depicting the alignment was produced, in which the target structure is shown in green, the structure determined without the dihedral restraints in magenta and the structure determined using the dihedral restraints in blue.

The first set of data (results shown in Figure 2.6) included [C'-H, N-H] from two alignment media. The resulting structure without the use of the dihedral restraints was 2.8Å from the x-ray structure whereas using the dihedral restraints resulted in a structure that was just 1.4Å away from the target.

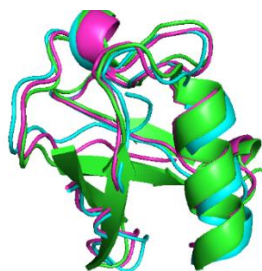


Figure 2.6. Alignment for the structure with dihedral restraints (magenta), without (blue) and x-ray structure of ubiquitin (green) for the first set of RDCs ([C'-H, N-H]x2).

The second set of data (results shown in Figure 2.7) included only [N-H] RDCs in two alignment media. The resulting structure without the dihedral restraints exhibited a bb-rmsd of 11.6Å whereas the structure that utilized dihedral restraints exhibited structural deviation of just 2.0Å.

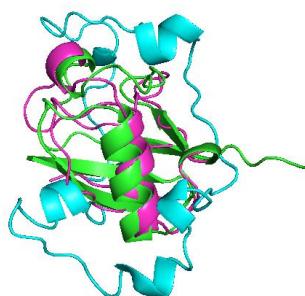


Figure 2.7 Alignment for the structure with dihedral restraints (magenta), without (blue) and x-ray structure of ubiquitin (green) for the second set of RDCs (N-Hx2).

The last set of data (results shown in Figure 2.8) included only [N-H] RDCs from just one alignment medium. The structure without dihedral restraints was over 21.1 Å away from the target structure whereas the structure calculation using dihedral restraints was just 2.4 Å away.

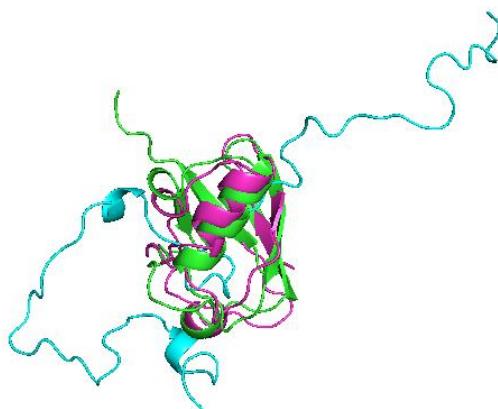


Figure 2.8. Alignment for the structure with dihedral restraints (magenta), without (blue) and x-ray structure of ubiquitin (green) for the second set of RDCs (N-Hx1).

Detailed results are shown in Table 2.1 for each of the datasets. The bb-rmsds in this table clearly show that the incorporation of data-driven dihedral increases the structure calculation ability of REDCRAFT. In addition, structural alignment of the three proteins from each set were aligned using a multiple structure alignment tool called MSTali[71]. Using the first set of RDCs, the three resulting structures retain 67 residues in common structurally. This indicates high level of structural similarity. However, when the dataset is reduced to the second and third set then the core residues in common drops to 29 and 22 respectively. This reinforces the structural dissimilarity of the structures in which the dihedral angles were used and those in which they were not.

Table 2.1. Results for each of the datasets is shown.

Set	RDCs (# Align Media)	BB-rmsd Without Dihedrals	BB-rmsd With Dihedrals
1	[C'-H, N-H] (2)	2.8	1.4
2	[N-H] (2)	11.6	2.0
3	[N-H] (1)	21.1	2.4



#### 2.1.4.5 Additional Scripts, Functionality, and Features

During structure calculation, thousands of different phi/psi combinations are explored. Currently, the REDCRAFT algorithm will automatically generate a .pdb file for the top structure as each amino acid is added to the structure. However, one may be interested in considering an ensemble of the top  $N$  structures, not just the “best” structure. To facilitate this analysis, pdbgen and pdbgen2 have been added which both generate .pdb files based on a string of phi/psi angles and a string of amino acids. Pdbgen can generate structures directly from the .out files that are created during a run of REDCRAFT and is able to read single character residue names. Pdbgen2, which does not require any options and only takes in a string of phi/psi angles and a string of amino acids as its arguments, is simpler to use and desirable for quick pdb construction. The pdbgen collection accommodates both basic and comprehensive structure generation from phi/psi angles. These programs can also function as standalone programs for quick pdb generation and verification where the other features of REDCRAFT are not necessary. The pdbgen tools will eventually make up part of the REDCRAFT GUI analysis suite where they can be better employed to help users find exactly where the intermediate protein structure may deviate during structure generation.

#### 2.1.5 Conclusions/Future Work

In this work, we have presented significant improvements to the REDCRAFT software package in the important areas of usability, accessibility, and core methodology. The inclusion of a GUI makes the software more usable by a wider audience. Incorporation of NEF standards makes the software compliant with a large suite of other widely available NMR software packages. In addition, the NEF import file allows for increased flexibility

of RDCs that can be utilized by REDCRAFT which will allow structure calculations of more complex and larger proteins, such as those that have been perdeuterated due to size. We have also shown that the improved decimation method allows the method to be used to calculate proteins that it was unable to complete before due to experimental noise. In addition, we presented incorporation of a dihedral restraint that was mined from the PDBMine database. Using these restraints, the structure of ubiquitin was characterized using just one RDC from one alignment medium. Structure calculation with so few RDCs per residue has, to date, never been achieved. Lastly, we introduced new standalone functionality to produce .pdb files from only phi/psi angles which is useful when analyzing ensembles of structures.

In future work, we plan to extend the REDCRAFT algorithm to also be capable of characterizing nucleic acids.

## 2.2 PDBMine: A Reformulation of the Protein Data Bank to Facilitate Structural Data Mining

### 2.2.1 Abstract

Large scale initiatives such as the Human Genome Project, Structural Genomics, and individual research teams have provided large deposits of genomic and proteomic data. The transfer of data to knowledge has become one of the existing challenges, which is a consequence of capturing data in databases that are optimally designed for archiving and not mining. In this research, we have targeted the Protein Databank (PDB) and demonstrated a transformation of its content, named PDBMine, that reduces storage space by an order of magnitude, and allows for powerful mining in relation to the topic of protein structure determination. We have demonstrated the utility of PDBMine in exploring the

prevalence of dimeric and trimeric amino acid sequences and provided a mechanism of predicting protein structure.

### 2.2.2 Introduction

Completion of the Human Genome Project in 1990[72] marked the beginning of the era of Big Data. Since then, various funding agencies initiated numerous large scale studies such as Structural Genomics Initiative[73], Protein Structure Initiative[1], and Genomic Data Commons[74] that have generated unimaginable volumes of data. Currently GO[75] houses over 7 million annotated genes and PDB[62] houses over 144,729 protein structures alone. Many other existing repositories of internationally collected data can be listed. While the advancement of technologies and scientific methods have contributed to the large growth in the data volume, velocity, and variety, the collected data has not had the anticipated impact in expansion of our knowledgebase. The limited impact of these databases is due to the fact that these repositories are optimized for data deposition but not for data mining. Recognizing this limitation, various funding agencies (including NSF and NIH) have declared new initiatives with the objective of transforming data to knowledge. Such transformation will require re-representation of data in such a manner that facilitates mining and knowledge discovery. Here we present the first instance of transformation of the Protein Databank (PDB) that allows for discovery of knowledge.

Proteins are a class of macromolecules that perform various functions in cells including structural support, enzymatic activities, cell signaling, and more. Proper regulation of proteins is crucial for life in all living organisms. Proteins are made up of smaller subunits called amino acids that are structurally defined by a series of angles called dihedrals. Large scale mining of this data can potentially lead to prediction of structure,

function, binding sites, and better understanding of evolutionary relationships between organisms[71, 76].

The primary database that houses structural information of proteins, PDB, contains all the necessary information needed to probe for structural insights but is not configured in such a way as to make the task straightforward. For example, there is currently no capability to easily extract dihedrals from these coordinates for a given sequence. To extract these angles for use in structure prediction algorithms, one would need to first perform a sequence search across all proteins, download all the hits, and finally write a script, or series of scripts, to extract the coordinates from the PDB files to perform the calculation of dihedrals. Each step of this process is difficult and/or time-consuming. Our new database, PDBMine, will alleviate these difficulties and make it easy to extract information such as dihedral angles quickly and accurately from raw PDB coordinates for use in a variety of applications. In addition to various observations resulted from PDBMine, we present our its preliminary application in prediction of protein structure.

### *2.2.3 Background and Methods*

#### *2.2.3.1 Existing Databases and Their Limitations*

Protein DataBank or PDB (<https://www.rcsb.org/>) is historically the oldest international repository of macromolecular structures dating back to as early as 1971. Currently, PDB houses the three-dimensional coordinates of 144,729 protein structures and provides an array of search mechanisms. However, the search mechanisms provided by PDB are aimed at navigating and retrieving the contents of the database. Therefore, there is little to no capability of querying at a more fine-grain level that will allow for knowledge discovery. Over the past decade, there have been several attempts[77, 78] that aimed at

creating a database for fine-grained mining of protein databases. In 2006, the DASSD[77] was created as a database that housed short protein fragments (sizes 1, 3, and 5 amino acids). Through now an inactive website, users were able to enter a sequence of amino acids to receive structural information regarding the middle residue of the query. In addition, DASSD would also provide a prediction related to the secondary structure formation of a given fragment. The limitation of the input size (1, 3 or 5 residues) made predictions of larger proteins implausible. Furthermore, as we demonstrate in the results, fragments of size 3 or 4 residues do not provide sufficiently converged results to meaningfully define a protein structure. Therefore, it is critical that a database to be capable of querying all fragment lengths including k-mers with  $k \geq 5$ . Protein Geometry Database (PGD)[78] is another attempt at creation of minable database of protein structures. This database extended the maximum fragment size to 10 amino acids and added additional search criteria such as R-factor and x-ray resolution. However, the current version of the database contains information for only 16,000 protein structures determined from only x-ray. While PGD may provide mining of “high-quality” structures, the limited number of protein structures and methods of characterization could lead to erroneous or biased results, especially when dealing with proteins that are inherently difficult for x-ray diffraction (such as membrane proteins and proteins that undergo dynamics). The database presented in this paper aims to overcome the limitations of these implementations and therefore create a more complete and encompassing platform for analyzing local and global protein geometries.

### 2.2.3.2 Transformation of PDB into PDBMine

Data and their corresponding databases can be formulated and transformed to reduce their space requirements, to increase data retrieval speed, to serve as more secure repository of data, to name a few example final objectives. In this work we transformed the existing PDB data with the primary goal of providing a more useful structural mining database. To that end, we extracted the following information from each protein structure: the mechanism of structure determination, the primary sequence of every protein, backbone torsion angles, hydrogen bonding, surface accessibility, as well as the three-dimensional coordinates of each atom. The derived information was then captured in PDBMine (simplified schema shown in Figure 2.9). The program DSSP[79] was used to convert each of the downloaded *pdb* files to their corresponding *dssp* files, which contained the surface accessibility, hydrogen bonding information, and  $\phi/\Psi$  angles for each residue. The contents of the DSSP file was then stored in the PDBMine database. MySQL was chosen as the platform for the PDBMine and a series of Python scripts were developed to parse and store the data into the database.

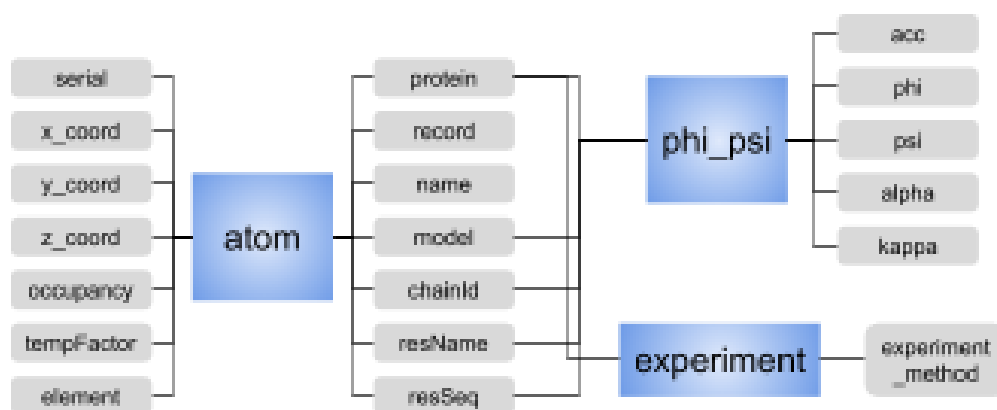


Figure 2.9. Database schema for PDBMine.

PDBMine consists of three tables: ATOM, PHI\_PSI and EXPERIMENT. The ATOM table contains the protein name, residue number, residue type, atom type as well as the atomic coordinates (X, Y, Z) for every atom in every protein. Storing the protein in this raw form allows for extracting spatial information such as “finding all the carbon atoms that are within 2Å of nitrogen” or “finding all the alanines that are within 5Å of a proline.” Such queries could be used to identify proteins with certain structural motifs. The PHI\_PSI table contains the dihedrals for all residues of each protein as well as other information collected from the DSSP program such as surface area accessibility. Finally, the EXPERIMENT table summarizes the metadata associated with the protein such as the experimental method of structure determination.

#### 2.2.3.3 Application of PDBMine in Structure Prediction

One application of the PDBMine database is prediction of backbone dihedral angles of a protein. Three dihedral angles phi ( $\phi$ ), psi ( $\Psi$ ), and omega ( $\Omega$ ) define the structural variability of the protein backbone at each amino acid. The collective effect of these torsion angles defines the overall structure of a protein. Due to the biophysical properties of the peptide bond, the  $\Omega$  torsion angle is generally fixed at  $180^\circ$  (or occasionally  $0^\circ$ ). Therefore, the effective degrees of freedom at any given amino acid are  $\phi$  and  $\Psi$ . Current methods of protein structure prediction and calculation[80, 81] rely on, at least to some extent, accurate prediction of dihedral angles for a given set of amino acids or a k-mer. These “local” dihedral predictions are used as scaffolding for the prediction of the full global structure. It, therefore, becomes important for the local k-mer geometries to be accurately predicted. If they are inaccurate then stitching the k-mers together to create the global structure will produce erroneous results.

Here we present an example application of PDBMine to facilitate more sophisticated and complete data analytics of the protein backbone dihedrals. In this application, we have created a frontend to the PDBMine with the specific task of collecting, analyzing, and reporting of the data. More specifically, this frontend accepts a protein primary sequence, and a search window size of  $k$  (where  $n \geq k$ ). The protein sequence is then automatically dissected into  $k$ -mers using a rolling window. Each  $k$ -mer is then queried in the database. The results for each of the queries are collected into a series of CSV files that contain the PDB accession number of the database hit, the chainID, model number, amino acid name, and the corresponding  $\phi/\psi$  angles. Furthermore, the backbone torsion angles for each residue is consolidated by combining the results for every one of the  $k$  places that the amino acid could appear in a  $k$ -mer rolling window. Finally, using the aggregated dihedrals for each residue and Kernel Density Estimation[38, 82], the most likely dihedral is predicted. All final and intermediate results are compiled and sent to the user via email.

#### 2.2.3.4 Evaluation Techniques

In 1963, G.N. Ramachandran noted the seminal observation that the  $\phi/\psi$  values in proteins adhere to a more restricted range of angles[56]. This restricted space of protein backbone dihedrals is denoted as Ramachandran plot (or R-Space). The first step in the evaluation of PDBMine was to query and recreate the previously reported R-Space[16, 46, 83] for each amino acid. The resulting distribution plots are then compared to the previously published and well accepted R-Spaces for each of the amino acids. This step will serve as a validation step through agreement with the previously reported work. As an extension, more complex and novel (previously unknown) R-Spaces were also created from extended  $k$ -mers (2-mers and 3-mers). The novel R-spaces serve as examples of new



information that can be produced from mining the PDBMine. Finally, we have demonstrated the potential of PDBMine in application to the challenging task of protein structure prediction. In this context, using the results from the database, the protein structure of ubiquitin was predicted purely based on statistical sampling of the backbone dihedrals. The predicted structure was compared to the x-ray structure currently published in PDB.

#### 2.2.4 Results and Discussion

##### 2.2.4.1 Database Creation

The total time consumed for downloading, parsing and uploading all proteins within the PDB was ~2016 hours, or 84 days. The final total space requirement for the database was 310 GB. This is an improvement over the space that is currently required (over 1 TB) to store the protein structures in pdb format in the PDB.

Of the 144,729 number of protein structures that were parsed, 3,764 of them required additional treatments due to file abnormalities such as the presence of DNA and RNA molecules, missing atoms, misnamed atoms, and others. These anomalies were addressed by designing and deploying specific scripts, after which, the final product was parsed and uploaded to PDBMine.

##### 2.2.4.2 Results of the Data Mining and Analysis

*2.2.4.2.1 Evaluation of Data* - As a prerequisite step, some basic analyses were performed to validate the content of PDBMine based on previously known information. The first of which was to calculate the abundance of each single amino acid and compare it to the statistics published from UniProt[84] (a database of all known protein sequences). Figure 2.10 shows a comparison of amino acid abundance from the two sources in a grouped bar

chart. The red and blue bars represent the calculated percentage occurrence of each amino acid in PDBMine and UniProt respectively. The two figures demonstrate very close agreement between two sources, indicating validity of the PDBMine’s data. In this figure, the largest observed difference is for the amino acid tyrosine (Y).

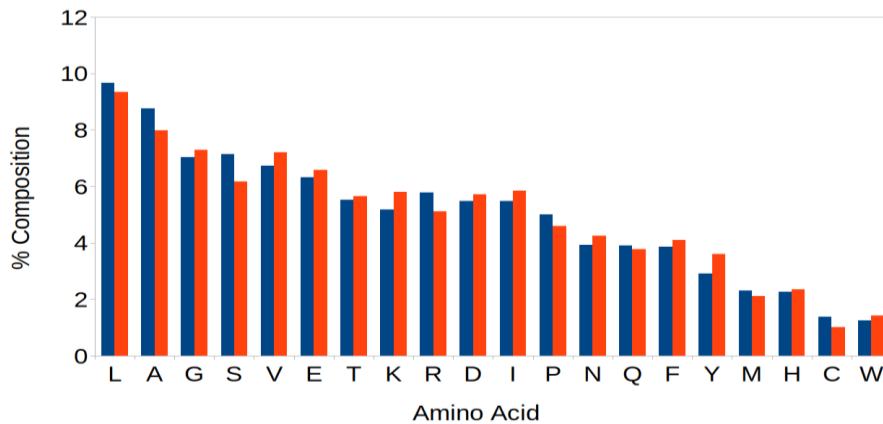


Figure 2.10. The abundance of each amino acid found in UniProt (blue) and PDBMine (red).

The abundance of amino acid 2-mers and 3-mers was also mined from PDBMine. Figure 2.11 shows the percentage appearance of all 2-mers (400 combinations) in all known protein structures. Although it is difficult to glean the exact count number for a given 2-mer from the figure, it is included here to show the general shape of the distribution. In particular, to illustrate that not all dimers are uniformly present. For instance, four dimers (LL, AL, AA, LA) occur ~700,000 times while three dimers (CW, CC, WC) only occur ~15,000 times (nearly 50 times less).

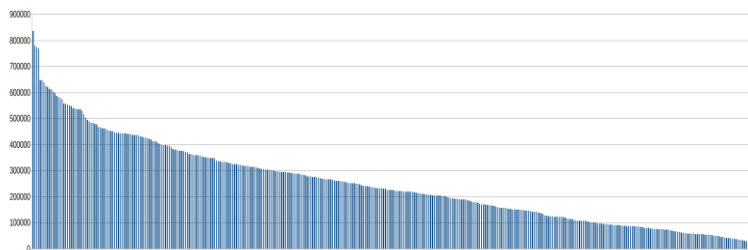


Figure 2.11. The general shape of the distribution of 2-mers in the database.

The percentage occurrence of 3-mers follows the same general shape as the 2-mer distribution shown in Figure 2.11. There are six 3-mers (ALL, EAL, ALA, AAL, LAA, AAA) that occur over 70,000 times within the database whereas there are eight (CHW, MWC, CMW, CCW, HWC, CIW, WWC, WCM) that occur less than 200 times. At the time of submission, there were no publications that detailed reasons as to the significance of some k-mers occurring more often than others.

In addition to comparing the distributions of amino acids, the R-Space was extracted for each of the 20 amino acids. These were visually compared to the known, accepted R-Spaces for single amino acids. For brevity, only the case of GLY and PRO are presented as they have R-Spaces that differ significantly from the other 18 amino acids. Figure 2.12a shows the comparison of the PDBMine generated (left) and accepted[85] (right) R-Spaces for GLY. Part b of Figure 2.12 depicts the same for PRO.

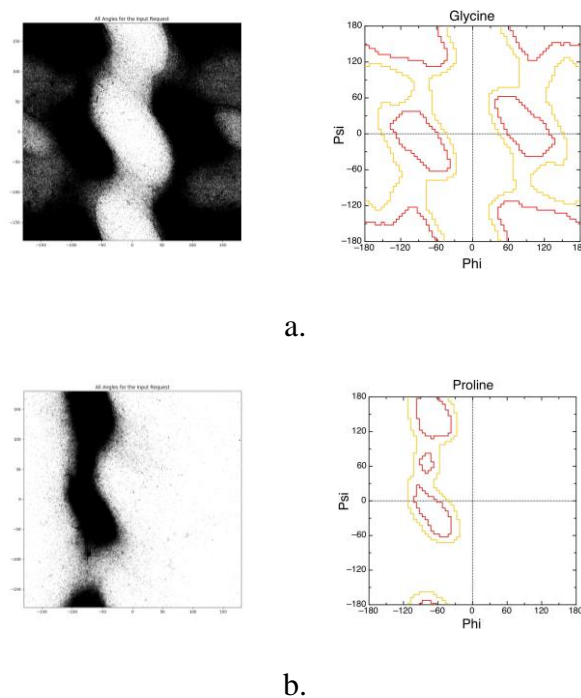


Figure 2.12. R-Space (PDBMine on left and accepted on right) for a) GLY and b) PRO.

*2.2.4.2.2 Prediction of  $\phi/\Psi$  Angles* - The utility of PDBMine can well exceed beyond the scope of a single amino acid. Figure 2.13a-d offer some useful insights for the common amino acid pair of glycine-proline. This motif occurs often in protein structures especially in the loop and turn regions. Figure 2.13a shows the R-Space for the glycine in the context of the glycine-proline combination. Proline is a relatively rigid amino acid whereas glycine is very flexible due to their respective sidechain configurations. In comparison to the typical glycine space (Figure 2.13a), Figure 2.13b shows a much more restrictive area of permissible torsion angles. Figure 2.13b depicts the R-Space for the proline of all glycine-proline amino acid pairs. In this case, the addition of the glycine does not change the R-space significantly for proline. Figure 2.13c-d show the R-Space for proline-proline pairs. In this pairing, the  $\phi/\Psi$  angles for the first proline (panel c) are significantly restricted compared to the traditional R-Space shown in Figure 2.13c. The second proline in the pair, shown in panel d, however, shows much better agreement with the traditional proline R-space. The proline-proline motif occurs in proteins fairly often with a current count size of 204,994 and, therefore, an increased understanding of its local structure will be of great benefit to computational methods.

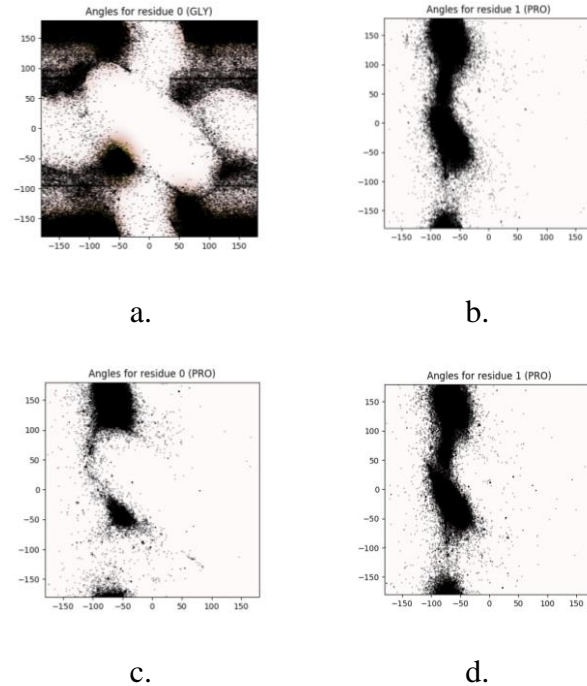


Figure 2.13 R-Spaces for a) GLY of GLY-PRO, b) PRO of GLY-PRO, c) first PRO of PRO-PRO and d) second PRO of PRO-PRO.

In addition to calculating the R-space for amino acids and 2-mers, this method can be extended to k-mer  $\phi/\psi$  prediction. One example is shown in Figure 2.14 depicting the R-Space for the 3-mer GLY-PRO-PRO. As it can be seen the space of allowed dihedral angles are significantly more limited compared to Ram-Space of a single amino acid (compare Figure 2.14a-c to Figure 2.12a,b).

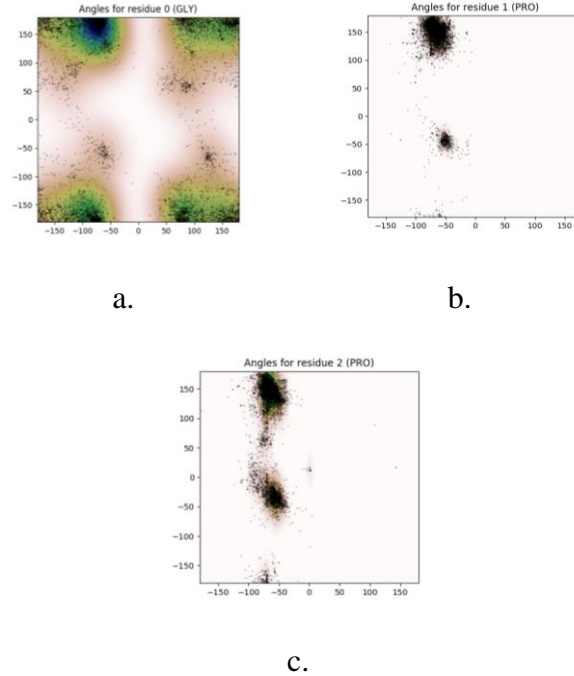


Figure 2.14. R-Space for the triplet GLY-PRO-PRO a) for GLY, b) for first PRO, c) for second PRO.

**2.2.4.2.3 Protein Structure Prediction** - The protein ubiquitin (76 residues) has been the subject of numerous studies by both experimental and computational methods[17, 46] of structure calculation. This makes it an ideal candidate for a proof-of-concept case. The dihedral angles of ubiquitin were calculated using a KDE-based prediction of k-mer dihedrals with k values of 3, 6, and 7. Examples of deviation in R-spaces for a given amino acid is shown in Figure 2.15. Notice that as the k increases (from left to right), the dihedral space becomes increasingly confined which leads to, as shown later, better structure prediction.

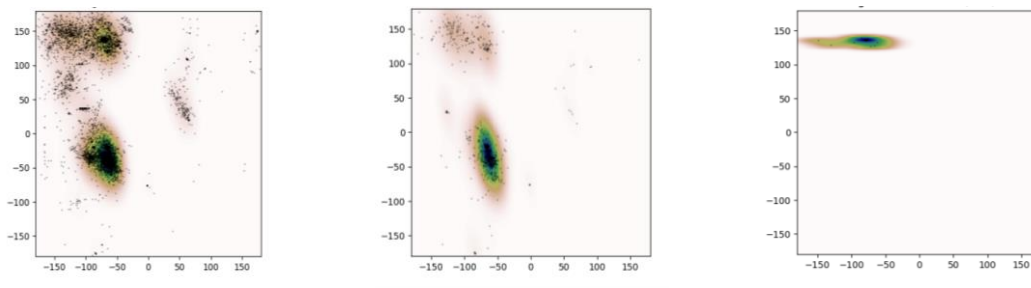


Figure 2.15. Examples of the differences for residue 14 of ubiquitin at lengths of  $k=3$ , 6 and 7

Structures were generated based on the results for the three different experiments (3-mer, 6-mer, 7-mer) using the program “pdbgen” included in the REDCRAFT[16, 20, 42, 86-89] software package. The structure using  $k=3$  (shown in Figure 2.16 in red) exhibited a backbone root mean squared deviation (bb-RMSD) of over  $22\text{\AA}$  to the crystal structure PDB-ID:1UBQ[90]. This indicates a low level of overall structural similarity. The resulting structures for 6 and 7 (shown in green and purple respectively in Figure 2.16) were similar with both exhibiting a bb-RMSD of around  $3.5\text{\AA}$  to the known crystal structure. This bb-RMSD indicates a reasonably high level of similarity between the two structures.

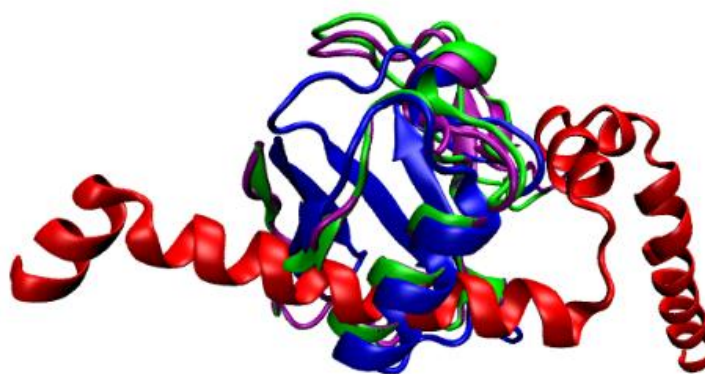


Figure 3.16. Resulting structures for  $k=3,6,7$  (red, green and purple) aligned to the x-ray structure (blue).

Further analysis using multiple structure alignment software msTALI[71, 76] showed that the conserved core between these three structures ( $k=6,7$  and the x-ray) contained 57

residues. The remaining 19 residues contributed to the divergence in structure (bb-RMSD). It is also worth noting, that a structural alignment including the result from k=3 yielded a core conserved region of 25 residues. While this indicates more regions of divergence, it also illustrates the amount of valuable information that is still present even at the k=3 level.

#### 2.2.4.3 Web Resources

A preliminary interface to the database has been created that will allow for fast, easy extraction of dihedral angles given a k-mer ([ifestos.cse.sc.edu/frontend](http://ifestos.cse.sc.edu/frontend)). The website was created using HTML/CSS, while the backend computation and data storage is accomplished using a combination of Python scripts and a mySQL database.

*2.2.4.3.1 Navigation* - The first page on the website the user inputs a window size, amino acid sequence, and an email address. After making a submission the user is provided with a summary of their input or an error message.

*2.2.4.3.2 Usage* - To submit a query, the user needs to provide a window size, an amino acid sequence, and a contact email address. The sequence can be submitted either as a single amino acid string (E N I E... etc.) or in a triplet format (GLU ASN ILE GLU... etc.). For the triplet format, the user needs to check the “Use AAA Formatting” option. Once submitted, the server will schedule the query and return the results once complete. These results contain a list of the predicted maximum likelihood angles for each residue in the request, plots of the KDE for each residue, and folder of the CSV files returned by the database. These CSV files contain a list of all proteins in the database that contained that k-mer along with the dihedral measures for each in which the user can perform their own additional analyses on the dihedral information.



*2.2.4.3.3 Capabilities* - Our local, fragment-based approach allows the user to obtain predicted structural information for sequences that have low global similarity with existing entries in the PDB. Changing the window size allows the user to control the amount of data returned. Larger sizes are more restrictive but will produce more well-defined results. Smaller sizes can be employed for sequences with unusually low representation in the PDB.

#### 2.2.4.4 Future Work

Future work will include improvements in two major areas: angle prediction and the web interface. Machine learning techniques such as traditional and deep neural networks can be used in place of the KDE method to improve the prediction of dihedral angles. Additions to the website will include advanced filtering including the ability to select only proteins characterized by certain experimental methods as well as the ability to select certain PDB ids to be excluded. In addition to these advances in capability, there will be additional graphical changes including onsite interactive visualization of R-Spaces for a given k-mer as well as automatic generation of protein structure ensembles from predicted angles.

#### 2.2.5 Conclusion

In this work, PDBMine, a database of dihedral angles mined from known protein structures, was presented. To demonstrate the validity of the data, known R-Spaces were generated and compared to their respective counterparts. In addition, preliminary results were shown for protein structure calculation using solely KDE-based prediction of dihedral angles. The web interface of this database allows for easy and efficient retrieval and analysis of dihedral angles for k-length amino acid sequences. The output of this website

can be easily incorporated into existing protein structure calculation tools for increased accuracy of models. In future work, more sophisticated mechanisms of prediction will be utilized, and improvements will be made to the web interface to allow for more flexible querying.

## 2.3 Structure Calculation and Reconstruction of Discrete State Dynamics from Residual Dipolar Couplings

### 2.3.1 Abstract

Residual Dipolar Couplings (RDCs) acquired by Nuclear Magnetic Resonance (NMR) spectroscopy are an indispensable source of information in investigation of molecular structures and dynamics. Here we present a comprehensive strategy for structure calculation and reconstruction of discrete state dynamics from RDC data that is based on the Singular Value Decomposition (SVD) method of order tensor estimation. In addition to structure determination, we provide a mechanism of producing an ensemble of conformations for the dynamical regions of a protein from RDC data. The developed methodology has been tested on simulated RDC data with  $\pm 1\text{Hz}$  of error from an 83 residue  $\alpha$  protein (PDBID 1A1Z), and a 213 residue  $\alpha/\beta$  protein DGCR8 (PDBID 2YT4). In nearly all instances, our method reproduced structure of the protein including the conformational ensemble to within less than  $2\text{\AA}$ . Based on our investigations, arc motions with more than  $30^\circ$  of rotation are identified as internal dynamics and are reconstructed with sufficient accuracy. Furthermore, states with relative occupancies above 20% are consistently recognized and reconstructed successfully. Arc motions with a magnitude of  $15^\circ$  or relative occupancy of less than 10% are consistently unrecognizable as dynamical regions within the context of  $\pm 1\text{Hz}$  of error.

### 2.3.2 Introduction

Structural elucidation, including the study of molecular complexes and characterization of internal dynamics of macromolecules, is often the requisite step in the molecular characterization of diseases. The breathing motion of myoglobin[91-94] can be cited as the historical instance of protein with internal dynamics that facilitates its biological function. Other proteins such as: hTS[95, 96], DHFR[3, 97, 98] and the carbohydrate recognition domain of Galectin-3[99] can be cited as other examples of dynamical proteins that are targets of pharmaceutical developments. Many RNA-binding proteins such as DGCR8, an integral component of the MicroRNA processing machinery[100], undergo conformational changes to enable the biological function of that protein. Therefore, development of methods leading to elucidation of structure and dynamics of proteins is of paramount importance. Study of dynamical proteins with X-ray crystallography is fundamentally difficult under the desiccated and restrictive crystalline environment that may interfere with the native-state dynamics of aqueous proteins. Nuclear Magnetic Resonance (NMR) spectroscopy, on the other hand, characterizes proteins in solution, which permits observation of conformational changes over different timescales. However, while NMR spectroscopy readily detects dynamical segments of proteins through the observation of T1/T2 relaxations[101-103] or Liparo-Szabo order parameters[104], such traditional approaches fail to provide atomic-level description of the conformational states. This is in part due to insensitivity of the commonly employed, short-range Nuclear Overhauser Effect (NOE) restraints to conformational changes. Furthermore, study of dynamics is complicated because functionally relevant events often take place on time-scales ( $\mu\text{sec}$ - $\text{msec}$ ) that are inherently difficult to observe by traditional

Liparo-Szabo approaches that are sensitive to time scales faster than the overall correlation time ( $\tau_c$ )[105-107].

The recent reintroduction of Residual Dipolar Couplings (RDC) acquired by NMR spectroscopy has presented new opportunities for structure determination and study of internal dynamics. Availability of RDCs has expanded the macromolecular investigations beyond structural characterization and into probing of internal dynamics[108] and molecular interaction. RDCs hold the promise to report on a large, comprehensive range of motional timescales spanning both sub- and supra- $\tau_c$  windows[109, 110]. The use of RDCs in study of dynamics fall into one of the two following categories: model-based and model-free approaches. The model-based approaches constitute some of the earliest approaches to investigation of internal dynamics. These methods utilize an existing protein structure (obtained by NMR spectroscopy or X-ray crystallography) and proceed by either assuming a fixed model of dynamics[5, 111] (typically a conical motion), or a presumed stochastic model[23, 112-114]. While these methods do not provide atomic level description of the conformational states, they can be used for quantitative analysis in amplitude of the internal dynamics. The model-free approaches[45, 50, 115-119] take advantage of the modern Molecular Dynamics Simulation (MDS) software such as CHARMM[120], NAMD[15], GROMACS[14], Amber[13, 121] or Xplor-NIH[122] to simulate the averaged observable RDC data over a course of conformational changes. These approaches can provide atomic-resolution conformational states, but at the same time rely on an existing protein structure as the starting point of the MD simulation. Independently, both approaches (model-based and model-free) proceed in two successive steps beginning with protein structures determined under the assumption of rigidity,

followed by characterization of dynamics. Although structure determination protocols based on the assumption of molecular rigidity may conveniently yield a structure, the degree of similarity between a static model of a protein structure and its many conformations remains poorly understood. Our recent work[20, 123] highlighted the possibility of obtaining erroneous structures for a protein that is undergoing internal dynamics. Consequently, mapping of dynamics onto a false static structure may lead to a compromised motional model. This can be attributed to the fact that it is conceptually difficult to separate structure from dynamics, because the two are intimately related. Thus, any attempt in structure elucidation that disregards the dynamics of the protein (or vice versa), may run the risk of producing faulty results. Furthermore, the strategy of structure-first followed by dynamics next, imposes collection of superfluous data, which may include: the traditional distance-based restraints and relaxation data to establish the existence of internal dynamics. Acquisition of the additional data inflates the cost and time requirements of these studies.

A conceptually attractive and alternative approach is to simultaneously characterize a protein's structure and its intrinsic dynamics[16, 17, 20, 41, 46, 124]. Ideally, such an approach could solely rely on RDC data carrying both, structural and dynamical information. However, the major bottleneck in utilization of RDC data in recent years has been attributed to a lack of RDC analysis tools capable of extracting the pertinent information embedded within this complex source of data. Nearly all legacy NMR data analysis software packages (i.e. Xplor-NIH, CNS[125], Cyana[126]) have been modified to accommodate RDC restraints. Other software packages have been developed in recent years specifically for structure calculation of macromolecules from RDC data[10, 16, 18,

23, 46, 47, 52]. Here we present a comprehensive approach for concurrent characterization of structure and dynamics of proteins from RDC data using the software package REDCRAFT[16, 20, 46]. Our approach permits structure calculation of proteins from a relatively sparse set of RDCs in the absence of dynamics. Here we extend our protocol to include identification and characterization of different modes of dynamics. Identification of the onset of dynamics and characterization of the mode of dynamics is based on the dynamic-profile analysis as implemented REDCRAFT. We demonstrate that discrete-state dynamical regions of a protein (when present) can be reconstructed based on perturbation of order tensors calculated from Singular Value Decomposition (SVD)-based[16, 127] mechanisms.

### 2.3.3 Methods

The presented methodology proceeds in four conceptual steps of: structure determination, identification of onset of dynamics, classification of the mode of dynamics, and reconstruction of different conformational states. Testing and validation reported in this work is based on simulated instances of dynamics and their corresponding RDC data. We have utilized REDCRAFT and a few other auxiliary programs to achieve our objectives. The following sections detail our methodology and approach in treatment of dynamics.

#### 2.3.3.1 Residual Dipolar Couplings

Residual Dipolar Couplings (RDCs) have been observed as early as 1963[57] and have been acquired for a number of structure determination studies including small molecules[128, 129], carbohydrates[130-133], nucleic acids[111, 134-137] and proteins[18, 49, 93, 138-141]. The RDC interaction phenomenon has been extensively

reviewed in the literature[7, 142, 143]. The physical principles[57, 101] that lead to manifestation of RDCs, and methods inducing alignment of biological macromolecules, have been fully described previously[142, 144-146]. Here we briefly review those components utilized by REDCRAFT. In addition, we limit our discussion to nuclei with spin quantum number of  $\frac{1}{2}$  and refer to the formula in Equation 2.3.1 from which all mathematical derivation of the RDC interactions (for a pair of spin  $\frac{1}{2}$  nuclei) begin. In this equation,  $\mu_0$  is the magnetic permeability of free space,  $\gamma_i$  and  $\gamma_j$  are gyromagnetic ratios of the interacting nuclei,  $h$  is Planck's constant,  $r$  is the distance separating nuclei  $i$  and  $j$ , and  $\theta$  is the angle between the magnetic field of the NMR spectrometer and a vector connecting atoms  $i$  and  $j$ .

$$D_{ij} = \frac{-\mu_0\gamma_i\gamma_j h}{(2\pi r)^3} \left\langle \frac{3\cos^2\theta_{ij}(t)-1}{2} \right\rangle \quad (2.3.3)$$

It is important to note that the RDC value  $D_{ij}$  (reported in units of Hz) is a function of the time-dependent angle  $\theta(t)$  averaged over time  $t$ , as represented by the angular brackets in Equation 2.3.1. This time averaging phenomenon may account for molecular motions caused by natural bond vibrations, internal dynamics, or overall tumbling of the molecule in the solution state. Mathematical transformation of Equation 2.3.1 can produce a computationally amiable formulation of the RDC phenomenon, as shown in Equation 2.3.2. In this representation of the RDC interaction,  $\nu$  signifies the normalized orientation of the interacting vector,  $s_{ij}$  denotes the  $ij^{\text{th}}$  element of the Saupe order tensor matrix,  $S_{ii}$  represents the principle order parameters, and  $\xi$  symbolizes the Eulerian rotation matrix that relates the molecular frame to the principal alignment frame. The remaining constants have been subsumed into a single constant,  $D_{max}$ .

$$D = D_{max} \bar{v}^T \cdot \begin{pmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} \end{pmatrix} \cdot \bar{v} = D_{max} \bar{v}^T \cdot \xi(\alpha, \beta, \gamma) \cdot \begin{pmatrix} S_{xx} & 0 & 0 \\ 0 & S_{yy} & 0 \\ 0 & 0 & S_{zz} \end{pmatrix} \cdot \xi^T(\alpha, \beta, \gamma) \cdot \bar{v} \quad (2.3.4)$$

Within recent years, various methods have been proposed[23, 69, 70, 82, 127, 147-152] to estimate the optimal order tensor from either a given set of RDCs, RDCs and a structure or just a structure. Each of these diverse approaches exhibit some advantages over the other existing methods but in practice the most accepted method of obtaining an order tensor is based on Singular Value Decomposition (SVD) analysis. SVD approach to calculation of order tensor is fast and provides the mathematically provable optimal solution. Despite its optimality, it poses certain challenges[45, 115, 119] when used in the context of studying dynamics from RDC data, which impose the use of structure-based methods of estimating order tensors[151]. Our presented methodology eliminates these challenges and resorts back to the use of more accepted SVD-based calculation of order tensors.

### 2.3.3.2 Categories of dynamics

To better facilitate the discussion of dynamics we enumerate three distinct dimensions of dynamics, namely: Temporal, Structural and Alignment as shown in Table 2.2. Along the Structural mode of dynamics, we define two categories of Rigid-body and Uncorrelated modes. Similar to previous definitions[12, 13, 122], Rigid-body dynamics is defined as dynamical regions that maintain a constant internal structure as a function of time, while the Uncorrelated dynamics is defined as alteration of structure as a function of time. Therefore, it is meaningful to describe the structure of a dynamical region if it is engaged in a Rigid-body dynamics, and not so for an Uncorrelated mode of dynamics. The



temporal dimension of dynamics can be defined by two categories of Discrete-state and Continuous-state dynamics. The distinction between the two is solely based on the temporal occupancy of conformational states that are visited during the trajectory of the dynamics. The Alignment mode of dynamics can be described by homogeneous and heterogeneous modes of alignment where the homogeneous mode of alignment assumes fixed alignment of the protein (within the same alignment medium) as a function of conformational changes. In contrast, in the heterogeneous mode of dynamics, alignment of a protein is altered as a function of the conformational changes. In principle all eight combined modes of dynamics should be possible with examples of all four combination of Structural and Temporal modes of dynamics having already been identified and presented in the literature[3, 96-98, 153]. In this report we investigate the combination of Rigid-body, Discrete-state dynamics with the explanation that it represents biologically most likely event, and that the remaining three modes (combinations of Structural and Temporal modes) can be approximated as a Rigid-body and Discrete-state dynamics in some favorable instances. The discussion related to the alignment mode of dynamics needs to be deferred for our future work as it is extensive and therefore distracting at this point. Therefore, in our simulations we assume a homogeneous alignment of the protein.

Table 2.2. Different modes of dynamics

Structural	Temporal	Alignment
Rigid-body	Continuous-state	Homogeneous
Uncorrelated	Discrete state-state	Heterogeneous

The foundation of the presented work is based on reconstructing the trajectory of dynamics using discrepancies of order tensors reported from the static and dynamic domains of a protein. Therefore, the first step in the study of dynamics is the mathematical formulation of effects of dynamics on order tensors. Equation 2.3.3 formulates changes in the observable order tensor (denoted as  $\hat{S}$ ) as a function of time (or dynamics). In this equation the variable  $j$  denotes the  $j^{\text{th}}$  alignment medium and integration is performed over the entire life of the dynamics. It can be argued that biological systems perform cyclical motions (returning to some original state), therefore the lifetime of a dynamic event can be treated as finite and periodical. Discrete approximation of the continuous function shown in Equation 2.3.3 can be developed as shown in Equation 2.3.4. In this formulation  $\delta t$  serves as the discrete time interval of the observation, which if selected appropriately can provide an accurate approximation of a temporally continuous motion. This equation can be further simplified based on relative occupancies in different states of the dynamics. This simplification can take place if the temporal occupancy of the conformational continuum is primarily in a small number of stable states (transient states are negligible). Under these conditions Equation 2.3.5 can be formulated and adopted in recovery of the primary conformational states of a Discrete-state dynamics. In this equation the entity  $S_j^i$  denotes the order tensor reported from the  $i^{\text{th}}$  conformational state within the  $j^{\text{th}}$  alignment medium where  $\rho_i$  is the relative occupancy of the  $i^{\text{th}}$  state. The second constraint shown in this equation enforces the fact that the sum of all relative occupancies should equate to 1 (or 100%).

$$\hat{S}_j = \int_{t=0}^{\infty} S_j(t) dt \quad (2.3.5)$$

$$\hat{S}_j = \sum_{k=1}^n S_j(k \cdot \delta t) \quad (2.3.6)$$

$$\begin{cases} \hat{S}_j = \sum_{i=1}^n \rho_i S_j^i \\ \text{Subject to: } \sum_{i=1}^n \rho_i = 1 \end{cases} \quad (2.3.7)$$

### 2.3.3.3 REDCRAFT

REDCRAFT[16, 17, 20, 46, 88] is designed for structure determination purely from orientational restraints. REDCRAFT deploys a powerful search mechanism that is significantly different from traditional optimization techniques, allowing for the same accuracy in recovery of structures compared to other algorithms while utilizing less data. The REDCRAFT software package has been previously described in depth and it is available for download from <http://ifestos.cse.sc.edu>. In this section we will present only the features that are relevant during the study of structure and dynamics of proteins.

REDCRAFT is well suited for the study of structure and dynamics because of its key feature of calculating the optimal structure by appending one residue at a time. This elongation process is consistent with the biological synthesis of proteins and allows for progressive examination of the rigidity assumption of a protein's structure. The averaging of order tensors due to internal dynamics leads to differences in the observed order tensors between the static and rigid components of a molecule. The differences of the order tensors result in an inherent inability to produce a structure that will consistently satisfy the orientational constraints between the static and dynamical regions. These inconsistencies can be identified from REDCRAFT's *dynamic-profile* that is produced during a structure calculation session. *Dynamic-profile* is formally presented and further discussed later.

The second feature of REDCRAFT that further enables study of structure and dynamics of proteins, emanates from its ability to conduct fragmented reconstruction of a protein. In general, structure of a given protein can be created in numerous fragments

because of data availability, biological importance, or study of dynamical regions that undergo Rigid-body dynamics. Study of *dynamic-profile* allows for identification of hinge regions, which can then be used to establish different dynamical domains of a protein for fragmented calculation of structures. The *dynamic-profile* has been described previously but to facilitate a better discussion, it is briefly discussed in section 0. Relevant to the current discussion, fragmented structure calculation that can be initiated based on analysis of a *dynamic-profile* allows structure reconstruction of all rigid components of a protein, although they may be dynamical with respect to each other. Once the individual structure of the rigid fragments within a protein are reconstructed, they can be assembled under a dynamics scheme that reconcile the differences in the observed order tensors across all alignment media.

#### 2.3.3.4 Dynamic Profile of REDCRAFT

The first step in investigating internal dynamics of a protein is to identify the hinge regions that give rise to the internal movement. It is also important to establish the structural mode of dynamics (Rigid-body versus Uncorrelated) after discovery of the onset of dynamics. Previously presented *dynamic-profile* that is produced by REDCRAFT can assist in discovery of the onset of dynamics and structural mode of dynamics. An example of a typical *dynamic-profile* for a static protein is shown in Figure 2.17. Under typical and non-anomalous conditions, a *dynamic-profile* will start with a very low RDC-rmsd score (due to initial lack of RDC data), will monotonically increase until arriving at a maximum value, followed by a final phase that is characterized by a plateauing of the RDC-rmsd score that is in agreement with the data acquisition error. Any significant departure from this typical profile is indicative of some anomalous conditions. The anomalous conditions

may consist of non-standard amino acid geometries (e.g. cis-Pro, impermissible dihedrals, non-standard bond lengths, etc.), existence of internal dynamics or mis-assignment of the RDCs, to name a few. Of particular interest to the discussion presented here, we will observe alternations of *dynamic-profile* as the means to identify the onset of dynamics and distinguish different structural modes of dynamics. Dynamic profiles can be generated for forward (N-terminus to C-terminus) or backward (C-terminus to N-terminus) analysis of a given protein. The forward and backward *dynamic-profiles* can help to corroborate the same anomalous regions with different degrees of certainty.

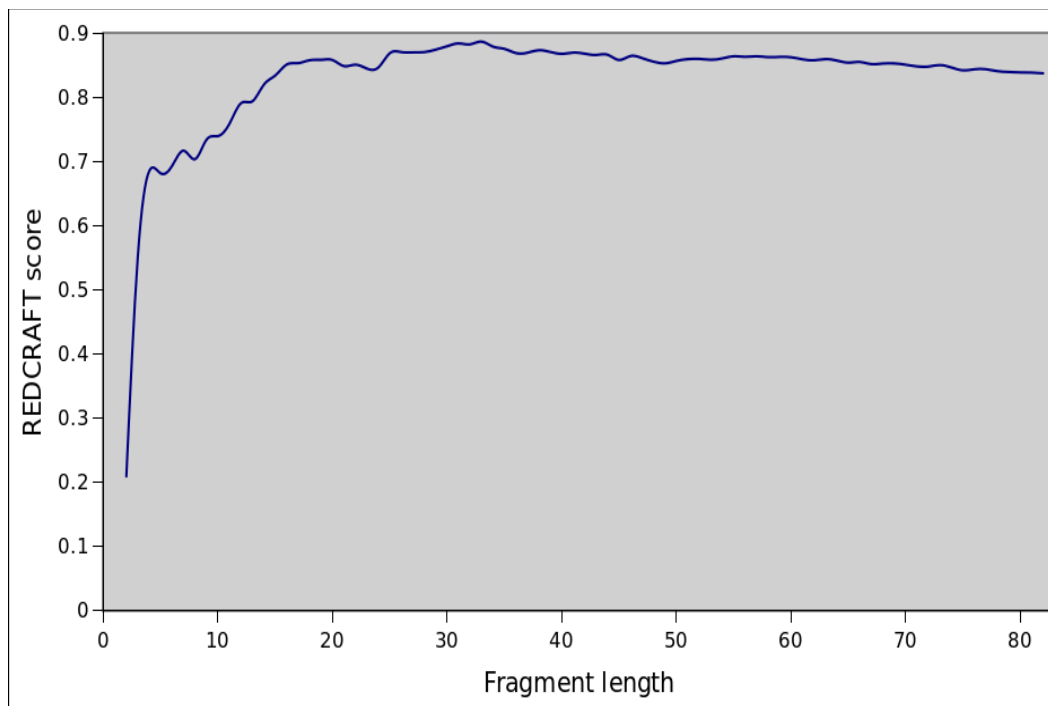


Figure 2.17. Example of a typical dynamic-profile for the protein 1A1Z in the absence of internal dynamics with simulated  $\pm 1\text{Hz}$  of uniformly distributed noise.

Analysis of REDCRAFT's *dynamic-profile* takes place in two steps. The first step serves to identify any form of structural anomalies by observing any deviation from a typical profile. The second step utilizes the ability of REDRAFT to perform a fragmented

structure determination of a protein. Once the point of anomaly is established, a new session of structure determination can be initiated a few residues in advance of the point of anomaly. The behavior of the *dynamic-profile* will be indicative of the structural mode of the dynamics. In section 0 we present results that demonstrate the use of this approach in discovery of onset and identification of structural mode of dynamics. Our exploration will consist of simulated Rigid-body dynamics and Uncorrelated dynamics using the protein 1A1Z as the target of our investigations. The specifics of the simulated dynamics are discussed in section 0.

#### 2.3.3.5 Theoretical treatment of dynamics

The following steps (also shown in Figure 2.18) describe our overall strategy in calculation of structure and characterization of dynamics:

1. Proceed in structure calculation with REDCRAFT under the assumption of structural rigidity.
2. Upon identification of internal dynamics from *dynamic-profile*, embark on fragmented study of dynamics for each region that exhibits internal structural rigidity.
3. After successful completion of fragmented structure calculation, establish the rigid and dynamical fragments through comparison of observed order tensors in all alignment media. Comparison of order tensors across different domains can establish static domains and dynamic domains. Fragments can be collected into relative rigid domains based on the similarity of their order tensors.
4. Construct models of dynamics that successfully explain the differences of the observed order tensors between the static and dynamic domains in all alignment media.

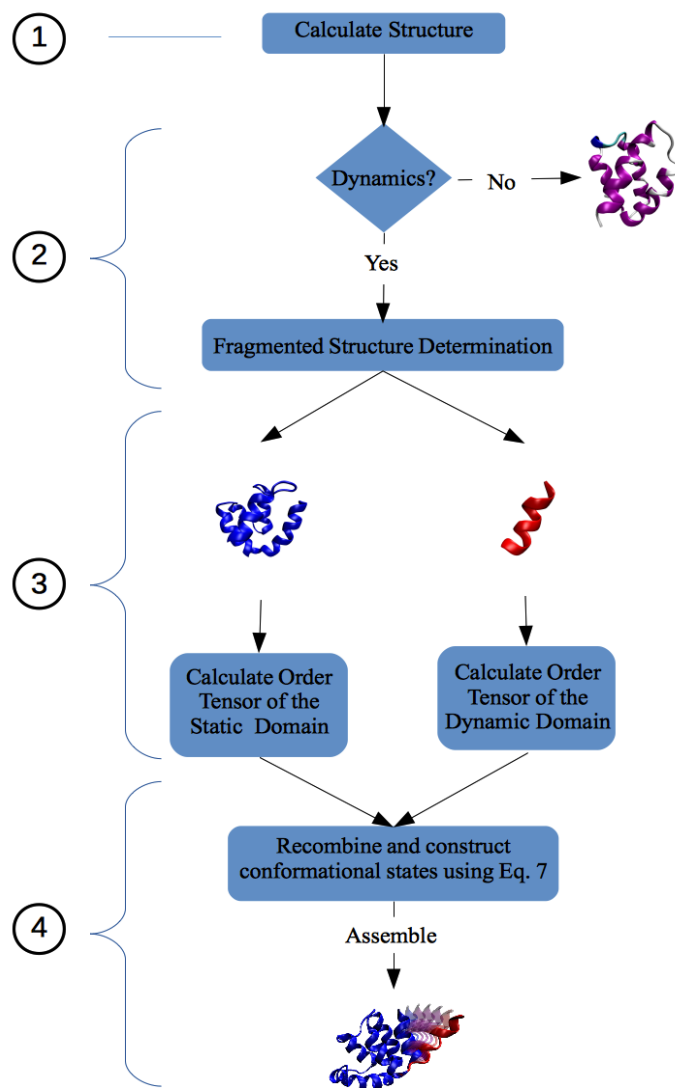


Figure 2.18. A diagram of illustrating our strategy for simultaneous characterization of structure and dynamics using REDCRAFT.

The scientific basis, technical requirements and procedures to establish steps 1-3 have been previously described and can easily be accomplished through the use of REDCRAFT and REDCAT software packages. However, additional theoretical formulations and procedural analyses are required for step 4. To facilitate the development

of procedures to accomplish the objectives in step 4, we first submit that in the case of a two domain dynamics, it is possible to designate one of the domains as the static domain and the other as the dynamic domain. Although at first the principle of relative motion may appear to introduce some ambiguity to this designation, the presence of a third entity (the external magnetic field) against which all tumbling, vibrational motions and internal dynamics are observed, disambiguates the designation. It is therefore possible to uniquely designate one domain as the dynamical and the other as the static domain by simply observing the General Degree of Order[5] (GDO) for each domain. Furthermore, Equation 2.3.5 can be used as the basis of expansion to accommodate reconstruction of the individual discrete states as shown in Equation 2.3.6. In this equation the term  $S^a_j$  denotes the anchor order tensor in alignment medium  $j$  and it signifies the order tensor that would have been observed if the dynamical domain was fixed and void of dynamics. The anchor order tensor can be obtained from the static domain of the protein (domain with the highest GDO). The term  $\xi_i$  represents the Eulerian transformation (with its three corresponding angular arguments) that maps the Rigid-body structure of the dynamical domain from any arbitrary molecular frame to the frame that defines the  $i^{th}$  state of dynamics. The average observable order tensor on the left-hand side of the equation can be obtained within REDCAT by analyzing the structure of the dynamical domain using the experimentally acquired RDCs. Equation 2.3.6 can be used to formulate the objective function shown in Equation 2.3.7, which can be used to obtain solutions for four unknowns (relative occupancy and three Euler angles) that define each state of a given discrete dynamics. In this equation the symbol  $\|.\|$  denotes magnitude of the difference-matrix by summing the square of its elements. This equation can be repeated for each alignment medium, which will contribute



five additional independent equations to the overall system of equations. In total, defining  $n$  discrete dynamical states will require  $4n-1$  (relative occupancy of the last state can always be computed by one minus the sum of all the other occupancies) degrees of freedom, while  $m$  alignment media will provide  $5m$  number of equations. Therefore, a viable solution can be obtained so long as the criterion shown in Equation 2.3.8 is satisfied. Note that an important fact in combining information across all alignment media is that relative occupancies and orientation of the dynamical domains with respect to the static domain remain unchanged across all alignment media. We have used least-square minimization[154, 155] routine available in Maple 14 software package to obtain the solution to the Equation 2.3.7.

$$\begin{cases} \hat{S}_j = \sum_{i=1}^n \rho_i S_j^i = \sum_{i=1}^n \rho_i \cdot \xi(\alpha_i, \beta_i, \gamma_i) \cdot S_j^a \cdot \xi'(\alpha_i, \beta_i, \gamma_i) \\ \text{Subject to: } \sum_{i=1}^n \rho_i = 1 \end{cases} \quad (2.3.8)$$

$$f(\rho_{1..n}, \alpha_{1..n}, \beta_{1..n}, \gamma_{1..n}) = \begin{cases} \sum_{j=1}^m \|\hat{S}_j - \sum_{i=1}^n \rho_i \cdot \xi(\alpha_i, \beta_i, \gamma_i) \cdot S_j^a \cdot \xi'(\alpha_i, \beta_i, \gamma_i)\| \\ \text{Subject to: } \sum_{i=1}^n \rho_i = 1 \end{cases} \quad (2.3.9)$$

$$5m \geq 4n - 1 \quad (2.3.10)$$

### 2.3.3.6 Testing and Validation

Our general testing and validation strategy have relied on the use of simulated RDC data. The overall process consists of generating average sets of RDC data from different models of dynamics, reconstructing fragmented structures based on steps 1-3 as listed in section 0, followed by reconstructing the dynamical states from the recovered Euler rotations (after solving Equation 2.3.7). Following reconstruction of the discrete states, validation is based on quantifying the backbone deviation between the reconstructed and

target states (described further in section 0). In our experiments we utilized synthetic data from an 83 residue FADD protein (PDB ID 1A1Z) and the 213-residue human DGCR8 core (PDB ID 2YT4). The use of simulated data during the early stages of method development is critical. The use of simulated data to test a new method of recovering dynamical states is a common practice and has certain advantages. Prior knowledge of the dynamics (ground-truth) allows for meaningful comparison of the recovered results to the known model of dynamics to establish the accuracy of the recovery method. Furthermore, simulated scenarios allow for systematic exploration in strengths and limitations of the presented methodology. In the following subsections the models of dynamics, summary of synthetic data and structure validation procedure used in our experiments are described in detail.

#### 2.3.3.6.1 Simulated data

Simulation of RDC values for an arbitrary pair of nuclei requires a-priori knowledge of an order tensor. A Saupe order tensor can be expressed via a  $3 \times 3$  matrix, or by providing principal order parameters  $S_{xx}$ ,  $S_{yy}$ , and  $S_{zz}$  and rotational Euler angles  $\alpha$ ,  $\beta$ , and  $\gamma$ . In this report we use the latter formulation of an order tensor. Using the atomic coordinates, order parameters and Euler angles, REDCAT was used to produce computed RDC values. We have utilized a number of order tensors in our investigations to passively observe the dependency of our method on order tensors. Tables 2.3-5 summarize the order tensors used for each of our models of dynamics.

Table 2.3. Order parameters used for the complex 2-state model of dynamics.

	$S_{xx}$	$S_{yy}$	$S_{zz}$	$\alpha$	$\beta$	$\gamma$
S1	$-3.00 \times 10^{-4}$	$-5.00 \times 10^{-4}$	$8.00 \times 10^{-4}$	$0^\circ$	$0^\circ$	$0^\circ$
S2	$2.00 \times 10^{-4}$	$5.00 \times 10^{-4}$	$-7.00 \times 10^{-4}$	$-40^\circ$	$-50^\circ$	$60^\circ$

Table 2.4. Order parameters used for the simulated 2-state arc motion and the simulated DGCR8 dynamics.

	$S_{xx}$	$S_{yy}$	$S_{zz}$	$\alpha$	$\beta$	$\gamma$
S1	$3.00 \times 10^{-4}$	$5.00 \times 10^{-4}$	$-8.00 \times 10^{-4}$	$0^\circ$	$0^\circ$	$0^\circ$
S2	$-4.00 \times 10^{-4}$	$-6.00 \times 10^{-4}$	$1.00 \times 10^{-3}$	$40^\circ$	$50^\circ$	$-60^\circ$

Table 2.5. Order parameters used for the complex 3-state model of dynamics.

	$S_{xx}$	$S_{yy}$	$S_{zz}$	$\alpha$	$\beta$	$\gamma$
S1	$3.00 \times 10^{-4}$	$5.00 \times 10^{-4}$	$-8.00 \times 10^{-4}$	$0^\circ$	$0^\circ$	$0^\circ$
S2	$2.00 \times 10^{-4}$	$5.00 \times 10^{-4}$	$-7.00 \times 10^{-4}$	$-40^\circ$	$-50^\circ$	$60^\circ$
S3	$-7.00 \times 10^{-4}$	$-1.00 \times 10^{-4}$	$8.00 \times 10^{-4}$	$20^\circ$	$-40^\circ$	$20^\circ$

Simulated RDC data may also be modified to include the addition of simulated error or noise. Unless specified otherwise, all simulated RDCs are accompanied by a uniform random change in the RDC values in the range of  $\pm 1$  Hz, and contain the following set of RDCs:  $[C'-N, N-H, C'-H, C_\alpha-H_\alpha]$ . To simulate different percentages of occupancies Equation 2.3.9 was used to average the sets of RDCs from different conformations, where  $\rho_i$  and  $RDC_j^i$  denote the relative occupancy and RDC values for vector  $j$  in the  $i^{th}$  conformational state respectively. In this equation  $n$  is the total number of discrete conformational states.

$$\begin{cases} \overline{RDC_j} = \sum_{i=1}^n \rho_i \cdot RDC_j^i \\ \text{Subject to: } \sum_{i=1}^n \rho_i = 1 \end{cases} \quad (2.3.11)$$

### 2.3.3.6.2 Simulated 2-state dynamics

Our exploration of 2-state dynamics consisted of two different models of dynamics. Both models were implemented on the FADD protein (PDB-ID 1A1Z) as an example of a helical protein. The helical nature of this protein presents unique challenges when studied by RDC data due to the parallel orientation of their *N-H* bonds. The two models of dynamics that have been included in this report consisted of an arc motion and a more complex motion resulted from rotation about two axes. The 2-state models of arc motion were explored by rotating the  $\varphi$  angle of the protein 1A1Z at the 71<sup>st</sup> residue (denoted by  $\varphi_{71}$ ) by 15°, 30° and 60°. Consequently, in the arc model of dynamics, this protein is segmented into two domains: a static domain that consists of residues 1-69 and the dynamic domain that consists of residues 73-83. An example of arc motion with 60° perturbation of  $\varphi_{71}$  is shown in Figure 2.19. In this figure the segment of the protein illustrated in blue is the static region while the red and green domains represent the two conformations of the dynamical region. It is noteworthy that this partitioning introduces additional challenges since the dynamical region is a single helix.

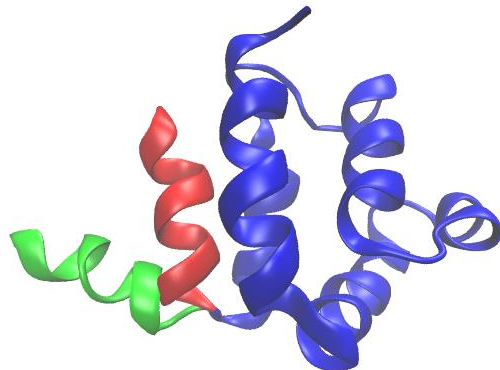


Figure 2.19. 2-state arc motion of the protein 1A1Z by 60° perturbation of the  $\varphi_{71}$  dihedral at residue 71.

The more complex motion (example shown in Figure 2.20) was created by performing a 30° rotation of the  $\phi$  and  $\psi$  angles at residue 58 (30° rotation of  $\phi_{58}$  followed by 30° rotation of the  $\psi_{58}$ ) of the protein 1A1Z. In this case the two domains were defined as residues 1-56 (the static region) and residues 60-83 (dynamic region). In Figure 2.20 the blue portion of the structure represents the static region, while the red and orange portions of the structure represent the two alternate states of the dynamical region.

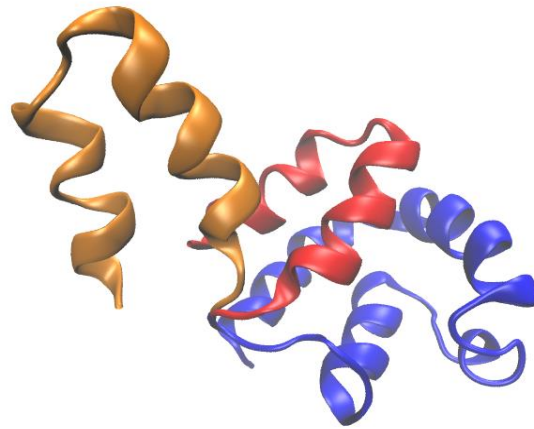


Figure 2.20. 2-state complex motion created by altering the dihedral angles of the protein 1A1Z at residue 58.

#### 2.3.3.6.3 Simulated 3-state dynamics

Our exploration of the 3-state dynamics consisted of building on the complex model of 2-state dynamics. Here the two states from the complex 2-state motion were used as states one and two of the complex 3-state motion. The third state was created by rotating the  $\phi$  angle of residue 58 (only  $\phi_{58}$ ) by 60° from the original structure. As in the case of the complex 2-state motion, the domains were defined by residues 1-56 and 60-83 as the static and dynamic domains respectively. The simulated three conformations are shown in Figure

2.21 where the red, green, and orange fragments illustrate states 1, 2, and 3 of the dynamical domain while the static domain is illustrated in blue.

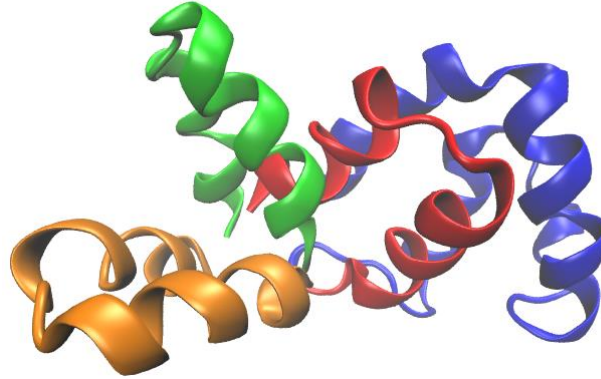


Figure 2.21. 3-state complex model of dynamics with blue representing the static domain and the dynamic domain shown in red, green and orange correspond to the conformational states 1, 2 and 3 respectively.

#### 2.3.3.6.4 Two-state jump model involving RBD2 of DGCR8

MicroRNA (miRNA) biogenesis follows a conserved succession of processing steps, beginning with the recognition and cropping of a miRNA-containing precursor (pre-miRNA) hairpin from a large primary miRNA transcript (pri-miRNA) by the Microprocessor. This Microprocessor consists of the nuclear RNase III Drosha and the double-stranded RNA-binding domain (dsRBD) protein DGCR8 (DiGeorge syndrome Critical Region protein 8) which is absent in individuals with DiGeorge syndrome. DGCR8 functions as an RNA-binding protein, yet the current crystal structure of the protein (PDB ID 2YT4) would require pronounced bending of the pro-miRNA substrate to engage both distant RNA-binding domains (RBDs). To address the biological implications of possible interdomain motions, we resorted to a molecular dynamic simulation of the DGCR8 protein to produce a more plausible RNA-binding model. Conceivably, a more likely scenario is

that DGCR8 adapts[156, 157] to allow for the RNA to bind by moving its two domains in tandem to create a favorable conformation to facilitate RNA binding. The motion between the two domains of DGCR8 is currently thought to be akin to a butterfly flapping its wings, with the linker region in between two RBDs as the mechanism of motion. To simulate this motion, rigid body dynamics was performed on 2YT4 using XPLOR-NIH. However, due to the dynamical nature of the protein, 2YT4 contained several gaps in various loop regions (residues 497-499, 584-591, 643-648, and 702-720) which impeded MD simulation. These gaps were remedied by the use of the I-TASSER[80, 158, 159] structure modeling tool. The resulting modeled structure exhibited  $0.5\text{\AA}$  of structural difference with respect to 2YT4 and contained no structural gaps. Using the complete structure, 50000 steps of rigid body dynamics were performed with step size of  $0.001\text{ psec}$  in a 400K bath temperature by keeping RBD 1 (residues 17-95) and RBD 2 (residues 126-203) rigid while permitting the linker region to fluctuate. The starting and ending frames of the trajectory were used as the two states with the RBD 1 of both frames superimposed to create a two-state jump motion for the RBD 2. The resulting two states are shown as the red and green structures in Figure 2.22 respectively. The orientational deviation between the two dynamical states was found to be  $5.4\text{\AA}$ . Average sets of RDC data were computed from the two conformations using the order tensors shown in Table 2.4 with  $\pm 0.5\text{Hz}$  uniformly distributed random noise added to the computed RDCs. These sets of RDCs were used for reconstruction of structures by REDCRAFT in a procedure highlighted in section 2.3.3.5.

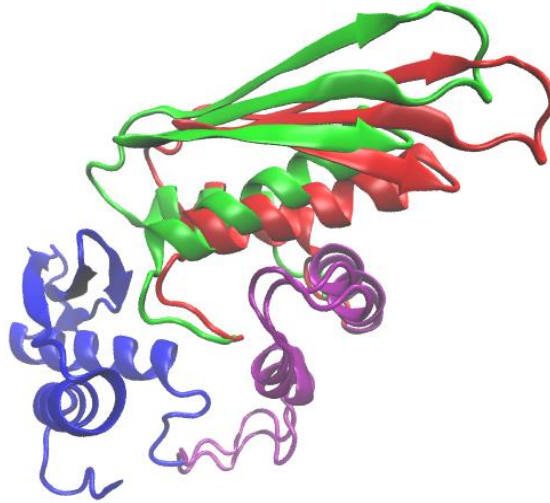


Figure 2.22. Two conformations of DGCR8 generated from Molecular Dynamics Simulation. RBD1(the static domain) is shown in blue, the linker region in purple and the two states of RBD2 are shown in red and green.

#### 2.3.3.6.5 Validation of results

Traditional method of reporting results for reconstructed structure of a protein is based on the measure of backbone-root-mean-square-deviation (bb-rmsd). In this application the simple use of bb-rmsd was not sufficient to report our findings since it would have generated results biased in favor of our method. Therefore, a more stringent approach was required in order to preserve relative orientation of a protein's fragments. Complete validation of the recovered structures in this report was comprised of three consecutive steps. The first step consisted of assembling the individual structural components including the different conformations of the dynamical region. Assembly of different conformational states was accomplished by utilizing the Euler angles that were obtained from minimization of the objective function shown in Equation 2.3.7. These Euler angles facilitated the correct orientation of the conformational domains with respect to the static domain. Furthermore, because information from more than one alignment medium



was used, any existing orientational degeneracies[160] (e.g. inversion degeneracy, etc.) were automatically resolved. It is important to note that upon the completion of this step, while the individual components of the protein were in correct orientational relationship with respect to each other, they may have exhibited a substantial translation in space.

During the second step of the validation, the target structure (including all of its conformational states) was rotated to a comparable orientation with respect to the reconstructed structure to serve as a template for measurement of the bb-rmsd similarity. During this step we have used MolMol[161] visualization software to optimally superimpose the static domain of the target protein onto the static domain of the reconstructed structure through rotational and translational modifications. Completion of this step provided a measure of backbone similarity between static domain of the target and reconstructed structures. The third step of our evaluation consisted of establishing the orientational accuracy of the reconstructed conformations for the dynamic domain by allowing only translational modifications (disallowing orientational modification) of the domains. Calculation of bb-rmsd based on optimized translation and disallowing rotational modification was performed by the software *backbone* that is included within the REDCRAFT software package. It is important to note that the reported bb-rmsd measures are an upper-bound estimates. It is beneficial to mention that the *backbone* software is capable of calculating bb-rmsd between two structures in three different ways: with no optimization (as is), translation optimized, translation and rotation optimized.

#### 2.3.4 Results and Discussion

In the following sections we provide results demonstrating the effectiveness of our approach in treatment of structure and dynamics of proteins. Our results first focus on the

ability of REDCRAFT to accurately identify the onset of dynamics and allude to the structural mode of the dynamic. Next, we present our results in reconstruction of conformations from two and three state dynamics. We conclude our results with a discussion of limitations of the presented work and anomalies related to the study of dynamics from RDC data.

#### 2.3.4.1 Discovery of onset of dynamics and structural mode of dynamics from dynamic-profile of REDCRAFT

As the first example in utility of the *dynamic-profile*, we present the case of 2-state dynamics. Here we utilized the dynamical model presented in section 2.3.3.6.2 (two states generated through perturbation of  $\varphi_{71}$ ) and utilized the averaged RDCs to perform a forward and reverse structure calculation of the protein 1A1Z. An example of the *dynamic-profile* of a 2-state dynamic can be seen in Figure 2.23(a). In this figure the blue and red profiles correspond to the forward and reverse structure calculations respectively. As it can be seen from Figure 2.23(a), and in contrast to the typical profile shown in Figure 2.17, an anomalous increase has manifested in the vicinity of residue 71 on both forward and reverse sessions of REDCRAFT. This is consistent with the model of dynamics that was used during this exercise. While both forward and reverse analyses exhibit an increase in the RDC score of the *dynamic-profile*, this phenomenon is more prominently observed in the case of the reverse structure determination than the forward. This inequality arises because in the case of forward run, the anomalous region is discovered after 73 residues and RDC data from only 11 residues exhibit inconsistencies with the remainder of the protein. This small portion will have a relatively smaller affect in perturbation of the RDC score reported by REDCRAFT. In contrast a much larger discrepancy is observed in the

case of reverse folding of the protein since a much larger portion of the data contributes to the observed inconsistencies.

A similar exercise was conducted for a 3-state dynamics (described in section 0) by altering the backbone dihedrals at the 58<sup>th</sup> residue. Figure 2.23(b) illustrates the *dynamic-profile* of this 3-state model of dynamics. Consistent with the model of dynamics, the *dynamic-profile* identifies the onset of dynamics at around residue 57-58. However, unlike the previous exercise and since a larger portion of the protein is undergoing dynamics, an approximately equal increase is observed in the *dynamic-profiles* of the forward and reverse structure calculation instances.

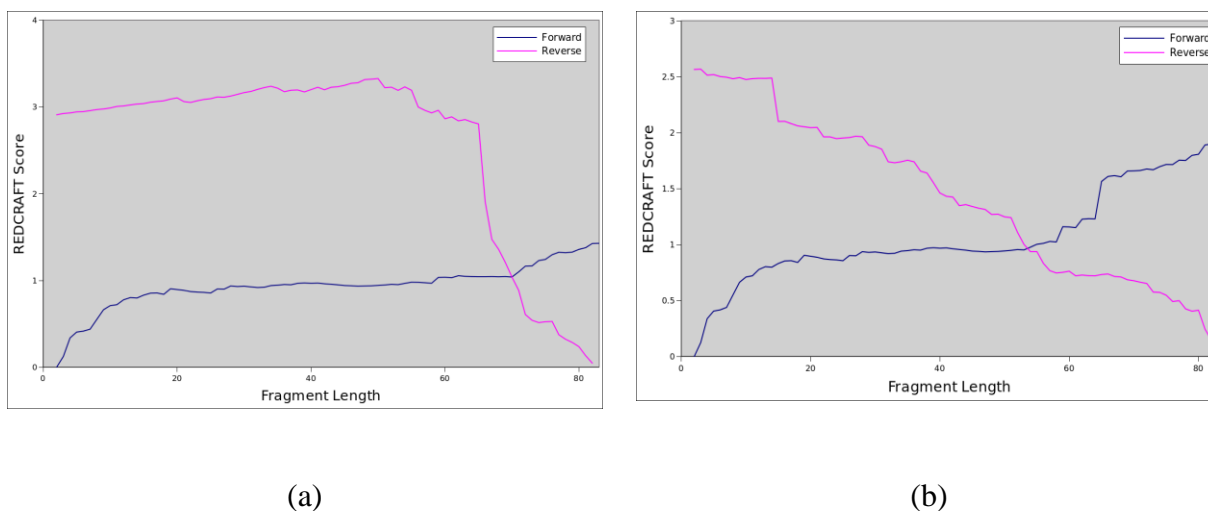
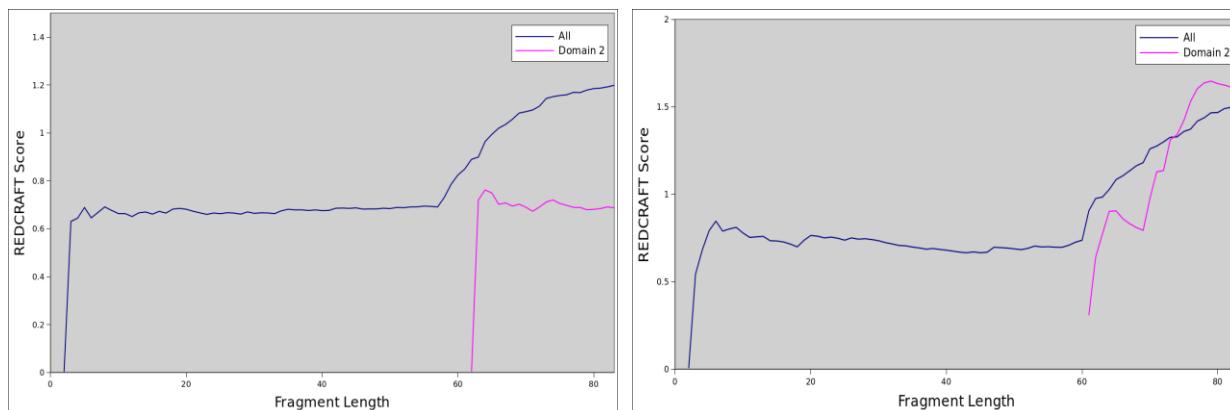


Figure 2.23. An example of the dynamic-profile for (a) 2-state model of dynamics and (b) 3-state model of dynamics

The above two examples demonstrate the ability of REDCRAFT in identifying the onset of internal dynamics via the use of *dynamic-profile* analysis. Structural mode of dynamics (Rigid-body versus Uncorrelated) can also be established by the use of fragmented study of a protein structure in REDCRAFT. In this context, structure calculation can be terminated prior to the onset of dynamics, and structure calculation of a

new fragment can be initiated a few residues past the onset of dynamics. Analysis of the *dynamic-profile* of the new fragment can help in establishing the structural mode of dynamics. The *dynamic-profile* of the new fragment undergoing Rigid-body dynamics will exhibit a typical pattern (similar to Figure 2.17) since it is internally rigid and consist of a structure that is internally static as a function of time. On the other hand, the Uncorrelated dynamics will exhibit a monotonically increasing score that indicates the lack of any consistent structure as a function of time. Figure 2.24(a) and Figure 2.24(b) illustrate examples of these two modes of dynamics simulated at  $\phi_{58}$  (in the case of rigid-body dynamics) and residues 58-83 (in the case of Uncorrelated dynamics). The *dynamic-profile* of the second fragment, for the case of Rigid-body dynamics that is shown in Figure 2.24(a), exhibits a normal behavior indicating successful reconstruction of a coherent structure. *Dynamic-profile* for the case of an Uncorrelated dynamics is shown in Figure 2.24(b) and it clearly exhibits a monotonically increasing behavior that indicates absence of a coherent structure. In the case of Rigid-body dynamics, upon recovering the structure of each domain, a measure of relative dynamics between the two domains can be established based on comparison of their corresponding order tensors.



(a)

(b)

Figure 2.24. Dynamic profile of a sample (a) Rigid-body dynamics and (b) Uncorrelated dynamics.

#### 2.3.4.2 2-state jump dynamics

As the first step in evaluating our approach to recovery of dynamical states, we resort to the arc motion of the 2-state dynamics described in Section 2.3.3.6.2. Here we explored alterations of  $\varphi_{71}$  by  $15^\circ$ ,  $30^\circ$  and  $60^\circ$  with different occupancies of the two states. We also evaluated our approach in recovery of 2-state dynamics of the complex motion (also described in section 2.3.3.6.2). Results of these experiments are shown in Tables 2.6 and Table 2.7 respectively. The results shown in these tables are segmented into sections corresponding to different states of occupancies. Our investigation in state occupancies starts with 50% occupancy of the first state and extends into 90%, in increments of 10%. The rows titled “Minimum” correspond to the lowest error in the minimized objective function (shown in Equation 2.3.7) in units of Hz corresponding to an  $N-H$  vector. A value approximately less than tenth of the experimental error (in this case 0.1Hz for 1Hz error) indicates successful reconstruction of the domains. In this table “Conformation” denotes the two states of dynamics, “BB-RMSD” corresponds to the backbone similarity of the

reconstructed states, and “Relative Occupancy” corresponds to the recovered occupancy of each state.

We begin the discussion of our results with the case of  $60^\circ$  arc motion (shown in Table 2.6). In general, all states of the dynamics were reconstructed very accurately including orientation of the states and relative occupancies. In some instances (such as 50/50) the relative occupancies were in error by as much as 13%. The only exercise that exhibited anomalous outcome was the case of 90/10. Here the first conformation was reconstructed with high degree of accuracy ( $0.37\text{\AA}$  with respect to the target protein) while the second state was created with bb-rmsd of  $9.4\text{\AA}$  with respect to its corresponding state. Our explanation for this behavior is that the low relative occupancy of this particular scenario marginalizes the perturbation of RDCs due to dynamics. The small perturbation of RDCs (in comparison to the  $\pm 1\text{Hz}$  noise) has therefore rendered its effect moot, thus collapsing the 2-state motion to a single rigid state. Since the effect of the second state is negligible, it was reconstructed in nearly an arbitrary orientation giving rise to its high bb-rmsd to the target conformation. This phenomenon is observed in other instances that are discussed in the future sections.

To further investigate the sensitivity of our method with respect to the magnitude of motion, the rotation of  $\varphi_{71}$  was reduced from  $60^\circ$  to  $30^\circ$ . The results of these experiments are shown in Table 2.7 and are very similar to that of the  $60^\circ$  dynamics except for the case of 80/20. In this case the second state was not reconstructed with sufficient accuracy. We speculate the reason for this inconsistency is the combination of a smaller angle of rotation and lower relative occupancy. It appears that at only a  $30^\circ$  rotation, a state with less than 20% occupancy may be subsumed into the original state as occurred in the case of 90/10

in the previous example. This explanation is further corroborated by the low relative occupancy of the second state and low objective function error.

To further investigate the effect and behavior of our approach on small and negligible motions, the case of  $15^\circ$  arc motion was examined. Although there existed internal variation of the second domain, the REDCRAFT dynamic profile does not identify internal dynamics, indicating the absence of any anomalous behavior from the perspective of the RDC data. Despite this finding we proceeded to reconstruct the two orientations and the results are shown in Table 2.6. The overarching observation that can be concluded from the results in this table is: accurate reconstruction of the first state, and nearly complete failure in reconstruction of the second state (both orientation and relative occupancy). In all of the cases highlighted in Table 2.6, the relative occupancy of the second state is very low with high objective function values (relative to the cases of  $30^\circ$  and  $60^\circ$ ). This is indicative of a negligible 2-state motion being subsumed into one rigid state, which is consistent with the results for  $60^\circ$  arc motion and 90/10 occupancy exercise. Both of these exercises help to establish the boundaries for the information content of the RDC data when simulated (or acquired) with  $\pm 1\text{Hz}$  of error. In summary, the particular instance of  $15^\circ$  motion did not provide sufficient alteration of RDCs (and therefore order tensors) to indicate the existence of internal dynamics at any relative occupancy.

Table 2.6. Results of 2-State 60°, 30° and 15° arc motion experiments.

Target Occupancies		60° Arc Motion		30° Arc Motion		15° Arc Motion	
50/50	Minimum	9.13×10 <sup>-11</sup> (0.23 Hz)		5.12×10 <sup>-11</sup> (0.17 Hz)		7.×10 <sup>-10</sup> (0.65 Hz)	
	Conformation	1	2	1	2	1	2
	BB-RMSD	0.93Å	1.02Å	0.46Å	0.44Å	0.78Å	7.8Å
	Relative Occupancy	0.63	0.37	0.45	0.55	0.96	0.04
60/40	Minimum	1.25×10 <sup>-10</sup> (0.27 Hz)		7.27×10 <sup>-11</sup> (0.2 Hz)		1.9×10 <sup>-10</sup> (0.34 Hz)	
	Conformation	1	2	1	2	1	2
	BB-RMSD	0.38Å	0.42Å	0.5Å	0.58Å	0.73Å	9.6Å
	Relative Occupancy	0.61	0.39	0.66	0.34	0.85	0.15
70/30	Minimum	1.31×10 <sup>-10</sup> (0.28Hz)		1.12×10 <sup>-10</sup> (0.26Hz)		3.48×10 <sup>-10</sup> (0.46Hz)	
	Conformation	1	2	1	2	1	2
	BB-RMSD	0.44Å	0.45Å	0.65Å	2.00Å	0.68Å	9.3Å
	Relative Occupancy	0.72	0.28	0.88	0.12	0.88	0.12
80/20	Minimum	2.37×10 <sup>-10</sup> (0.37 Hz)		9.4×10 <sup>-11</sup> (0.23 Hz)		7.13×10 <sup>-10</sup> (0.65 Hz)	
	Conformation	1	2	1	2	1	2
	BB-RMSD	0.59Å	1.52Å	0.66Å	5.2Å	0.66Å	9.1Å
	Relative Occupancy	0.85	0.15	0.95	0.05	0.88	0.12
90/10	Minimum	2.29×10 <sup>-10</sup> (0.37 Hz)		4.49×10 <sup>-11</sup> (0.16 Hz)		1.14×10 <sup>-9</sup> (0.82Hz)	
	Conformation	1	2	1	2	1	2
	BB-RMSD	0.37Å	9.4Å	0.5Å	6.9Å	0.58Å	5.2Å
	Relative Occupancy	0.896	0.103	0.98	0.02	0.95	0.05



The results from the complex 2-state model are shown in Table 2.7 and convey an outcome consistent with the case of arc motion. Both conformations were reconstructed with high degree of accuracy despite the complexity of the dynamics. However, it can be seen that as the relative occupancy of the second state decreases, accuracy of its reconstructed orientation deteriorates. This deterioration in performance is observable in the case of 80/20 and clearly so in the case of 90/10. In both cases the first state was reconstructed with reasonable accuracy while the reconstruction of the second state deteriorated as a function of occupancies. A relative occupancy of 10% can be seen as almost negligible in the course of a dynamic movement when using RDCs with  $\pm 1$ Hz of error.

Table 2.7. Results for 2-state complex dynamics experiments.

50/50	Minimum	$2.27 \times 10^{-10}$ (0.36 Hz)	
	Conformation	1	2
	BB-RMSD	0.76Å	0.83Å
	Relative Occupancy	0.42	0.58
60/40	Minimum	$1.6 \times 10^{-10}$ (0.31 Hz)	
	Conformation	1	2
	BB-RMSD	1.1Å	1.4Å
	Relative Occupancy	0.47	0.53
70/30	Minimum	$1.4 \times 10^{-10}$ (0.29 Hz)	
	Conformation	1	2
	BB-RMSD	1.2Å	1.6Å
	Relative Occupancy	0.53	0.47
80/20	Minimum	$6.04 \times 10^{-11}$ (0.19 Hz)	
	Conformation	1	2
	BB-RMSD	0.69Å	2.3Å

	Relative Occupancy	0.66	0.34
90/10	Minimum	$1.7 \times 10^{-10}$ (0.32 Hz)	
	Conformation	1	2
	BB-RMSD	0.83Å	6.33Å
	Relative Occupancy	0.95	0.05

In summary, results in Table 2.6 seem to indicate that at just 15° of movement, our approach is able to reconstruct one of the states (State 1) with reasonable accuracy, but fails to reconstruct the second state. However, it can be observed that when the motion is extended to a 60° or 30° movement (results shown in Table 2.6), then both states can be reconstructed with reasonable accuracy so long as the relative occupancies exceed 20%. The general explanation for both cases is that the contribution of dynamics is less than the experimental noise, and therefore meaningful calculations are moot.

#### 2.3.4.3 3-state jump dynamics

Results of the 3-state complex dynamics are shown in Table 2.8 for a number of different relative occupancies. Similar to the case of 2-state, the relative occupancies of each exercise are listed on the first column of this table. The “Minimum” value corresponds to the lowest value (in units of Hz scaled to N-H vectors) obtained from minimizing the objective function shown in Equation 2.8. This value helps to establish the success of the general approach; minimum values in the vicinity of tenth of the experimental noise indicate successful reconstruction of the states.

Table 2.8. Results for 3-state dynamics experiments.

50/25/25	Minimum	$2.9 \times 10^{-11}$ (0.13 Hz)		
	Conformation	1	2	3
	BB-RMSD	0.95Å	1.9Å	0.67Å
	Relative Occupancy	0.42	0.32	0.26
34/33/33	Minimum	$2.6 \times 10^{-11}$ (0.12 Hz)		
	Conformation	1	2	3
	BB-RMSD	1.4Å	0.38Å	1.3Å
	Relative Occupancy	0.25	0.41	0.34
50/30/20	Minimum	$3.4 \times 10^{-11}$ (0.14 Hz)		
	Conformation	1	2	3
	BB-RMSD	1.08Å	1.5Å	0.4Å
	Relative Occupancy	0.32	0.34	0.34
60/30/10	Minimum	$7.8 \times 10^{-11}$ (0.21 Hz)		
	Conformation	1	2	3
	BB-RMSD	0.64Å	1.3Å	1.3Å
	Relative Occupancy	0.55	0.35	0.1

As seen in Table 2.8, our presented method has successfully reconstructed the conformational states and rates of occupancies with less than 2Å in structural resolution. We note that there is a higher variability in the recovered measure of relative occupancies; variations as much as 0.19%.

#### 2.3.4.4 Recovery of DGCR8 discrete state dynamics

When combined with an appropriate analysis tool, RDC data can provide a significant reduction in data requirements. As an example, we present results for the analysis of structure and reconstruction of a simulated 2-state model of dynamics by using

only a fraction of the entire structure. In this exercise the structure calculation of the protein 2YT4 was limited to only a portion of the static and dynamic domains. More specifically, residues 17-41 were used as representative of the static domain, and residues 130-142 represented the dynamical region. Here we can establish the relative dynamics of two domains that are separated from each other (in space and sequence) without the need to study the entire protein. This exercise also helps to gain some insight as to the size of a fragment that is needed for successful reconstruction of dynamics. Each of the domains were reconstructed with bb-rmsd of 1.5Å to their corresponding portion of the target protein respectively. Analysis of the order tensors strongly supported the existence of internal dynamics and results of our conformational reconstruction are summarized in Table 2.9. Based on results shown in this table, the two states of dynamics were reconstructed with reasonable degree of accuracy. Figure 2.25 provides an illustration of these results for the two recovered states of DGCR8.

Table 2.9. Results in recovery of DGCR8 discrete state dynamics.

50/50	Minimum	2.2×10 <sup>-10</sup> (0.36 Hz)	
	Conformation #	1	2
	BB-RMSD	1.5Å	1.5Å
	Rate of Occupancy	0.54	0.46

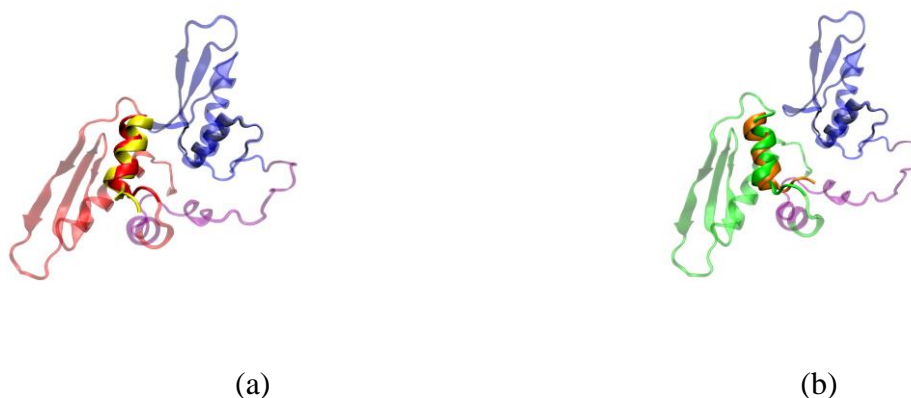


Figure 2.25. Reconstructed states from the DGCR8 experiment. Opaque renderings denote the fragments of the protein reconstructed. (a) The first conformation of the target structure is shown in red and the corresponding conformation is shown in yellow. (b) The second conformation of the target protein is shown in green and the corresponding reconstructed conformation is shown in orange.

#### 2.3.4.5 Modeling of 2-state dynamics as 3-state dynamics or a 3-state as 2-state

Results presented in the previous sections assume a-priori knowledge in the number of dynamical states. It is reasonable to consider the cases where the number of stable conformations is not known prior to analysis. In this case, a parsimonious approach can be employed to assist in the discovery of the appropriate jump states. More specifically, a 2-state dynamics can serve as a starting point of any investigation. Total number of conformational states can be explored incrementally until a satisfactory result is achieved. A satisfactory result is quantified by fitness of experimental data to the computed ones to within the data acquisition error. To demonstrate this approach, analysis of 3-state dynamics described in section 0 was utilized. Based on this parsimonious approach, the reconstruction of conformations will proceed based on the assumption of a 2-state dynamics. Results of the 2-state recovery of the 3-state dynamics are shown in Table 2.10. In principle, and in agreement with the results shown in this table, the incomplete modeling should be problematic and manifest itself in an unacceptably high objective function value.

In Table 2.10, the left-most column indicates the true relative occupancies of each state during the simulation of dynamics. The information marked as “Minimum” denotes fitness of the objective function (Equation 2.3.7) scaled to the units of Hz for  $N-H$  vectors. Increasing the number of states to 3, produces the results shown in Table 2.10 with minimum values of the objective function that clearly indicate successful recovery of states. The cases of 60/30/10 and 50/30/20 exhibited potentially acceptable objective functions because they can be treated as a two state dynamics by disregarding the state with relative low occupancies (10% or 20%). This serves as another affirmation that relative occupancies of less than 20% are potentially negligible within the framework of  $\pm 1$ Hz of experimental error.

Table 2.10. Results for modeling of a 3-state dynamic as a 2-state.

True Occupancies	Minimum
34/33/33	$2.99 \times 10^{-8}$ (4.21 Hz)
50/25/25	$4.097 \times 10^{-7}$ (15.56 Hz)
50/30/20	$5.8 \times 10^{-9}$ (1.85 Hz)
60/30/10	$6.9 \times 10^{-9}$ (2.02 Hz)

Conversely, a 2-state dynamics can be forced to be modeled as a 3-state. In theory, a 2-state dynamics should be classified as a 3-state dynamic where two of the recovered states correspond to the two conformations, and a phantom third state with a relative occupancy of 0%. To illustrate this point two experiments in which 2-state models of dynamics were forced into a 3-state recovery. Recovery of 3-state dynamics requires RDC data from at least three alignment media. The three alignment media shown in Table 2.5 along with the 2-state arc motion and 2-state complex motion described in Section 2.3.3.6.2

were utilized in this exercise. In both cases equal 50% relative occupancies were used to simulate the RDC data.

Table 2.11. Results for simulating 2-state dynamics in our 3-state dynamic eq.

Arc Motion (50/50/0)	Minimum	$3.15 \times 10^{-13}$ (0.013 Hz)		
	Conformation	1	2	3
	BB-RMSD	0.7Å	0.63Å	4-7Å
	Rate of Occupancy	0.47	0.50	0.03
Complex Motion (50/50/0)	Minimum	$1.6 \times 10^{-10}$ (0.31 Hz)		
	Conformation	1	2	3
	BB-RMSD	0.66Å	0.6Å	9-10Å
	Rate of Occupancy	0.44	0.55	0.01

As can be seen from Table 2.11, Conformation 3 in both the arc and the complex motions have occupancy rates of 0.03 and 0.01 respectively. These conformations correspond to neither state 1 nor state 2 of their respective model of dynamics. An occupancy rate of 1-3% is in practice negligible, making the corresponding state clearly inconsequential. Figure 2.26 shows the results from the 2-state arc motion with the extraneous conformation shown in yellow along with the two conformations (1 and 2 in Table 2.11) that align well with the original model of dynamics.

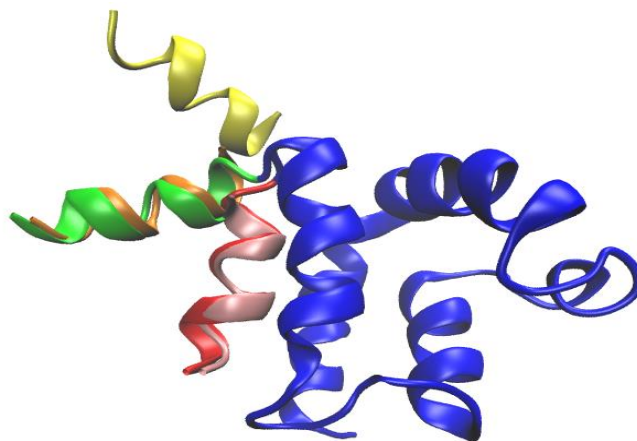


Figure 2.26. The resulting conformations from forced modeling of a 2-state dynamic as a 3-state are shown here. Fragments shown in red and green correspond to the two actual conformational states while yellow depicting the phantom irrelevant conformation with 1% relative occupancy.

The results of these experiments are important in the sense that they reveal interesting insights into the presented method. They show that inclusion of more data improves the preciseness in the reconstruction of domains as evidenced by the low bb-rmsd of the reconstructed states in Table 2.11. In addition, it can be reasonably argued that in application to natural examples of dynamics where the true number of discrete states are not known, our method appears to successfully identify the proper number of states that describes a model of dynamics.

#### 2.3.4.6 Limitations in recovery of discrete state dynamics

In section 0 the inability of the presented work to reconstruct conformations with relative occupancies less than 10% was demonstrated. In addition, the limitation in reconstructing conformational changes that are imposed by as small as  $15^\circ$  of arc motion was also demonstrated. Both of these limitations are due to the overall contribution of dynamics (either relative occupancy, or small motion) relative to the experimental



precision of data acquisition. Therefore, in the absence of any other information, these types of limitations diminish the efficacy of investigating dynamics from RDC data.

Approaches that rely on analysis of order tensors to recover conformational information are inherently limited by an upper-bound in the number of recoverable states. More specifically, these limitations arise from the fact that order tensors span a five dimensional space (degrees of freedom of an order tensor) and therefore regardless of the number of alignment media explored, no more than five independent alignment tensors can be obtained[162]. Considering the relationship shown in Equation 2.3.8, this imposes a limitation on our approach of recovering a maximum of six conformations.

Finally, some mechanisms of dynamics are more pathological than others. There are instances of dynamics that diminish and limit the information content of RDC data and therefore impede the process of structure reconstruction. For example, previously described effects of a C3 motion (three conformational states related by 120° rotations about a common axis) will converge the averaged anisotropic alignment (of any alignment medium) to an axially symmetrical (rhombicity of 0) order tensor with the primary orientation along the axis of motion. This convergence of order tensors may impose some degeneracies, which can cause challenges during the structure determination by RDC data.

### 2.3.5 Conclusions

RDCs can be an indispensable source of information in exploration of structure and dynamics of macromolecules. Here, we demonstrate that conformational changes of internal domains can be reconstructed to within atomic resolution. In addition, the relative occupancies of each state can be assessed with sufficient accuracy. We show successful applications to reconstruct dynamics consisting of a conical rotation, rotation about two

axes, and more realistic motions derived from an MD simulation. While we reliably reconstructed dynamics involving two and three states with high accuracy, our investigations revealed some inherent limitations associated with the extraction of dynamics from RDC data. In general, dynamics of small magnitude or states with low occupancies which cause perturbations of RDCs comparable to experimental noise cannot be reconstructed in a meaningful way. Typically, states within conformational ensembles populated less than 10% do not sufficiently perturb the corresponding RDCs. Similarly, arc motions smaller than  $15^\circ$  produce insufficient changes of RDCs and prohibit a meaningful investigation of dynamics. However, further studies are needed to describe and exploit the non-linear nature of RDC phenomenon. The inherent non-linearity of RDCs may profoundly influence the results. For example, the same magnitude of rotation about a sensitive axis (e.g. magic angle) may produce far larger perturbation of RDCs than some other insensitive axis of rotation.

Degeneracies associated with C3 dynamics (a scenario involving three equally populated states with C3 symmetry) that potentially impede RDC-based studies of structure and dynamics are particularly interesting. In such instances the anisotropy imposed by two independent alignment media converge to two very similar and symmetrical order tensors, thereby effectively reducing the accessible information to a single symmetric order tensor. Such unfavorable conditions inevitably interfere with studies of structure and dynamics and require comprehensive mathematical and theoretical treatments that are beyond the scope of this study in order to be rectified. However, such anomalies primarily occur within simple models of dynamics (such as arc motions) and seem to be absent in more complex cases.

In the presence of sufficient amount of RDC data, the presented method can be extended to describe conformational ensembles characterized by as many as six states. The maximum limit of six states is fundamentally imposed by the dimensionality of order tensors. Our future efforts will focus on better understanding degeneracies that may be encountered during the study of dynamics and will aim to reassess the problem independent of order tensors.

Finally, we reported outcomes of our approach in scenarios where the number of discrete conformational states is either under- or overestimated. When encountering an underestimated number of conformational states, the unacceptably high fitness of the objective function from Equation 2.3.7 will serve as a clear indicator. This mechanism will allow reexamination of the problem with an adapted number of discrete states. In the case of overestimated number of states within the ensemble, we have demonstrated that the conformational states are correctly reconstructed with an additional phantom state that is identified by a low relative occupancies. Such phantom states can simply be disregarded, and the primary states utilized in a simplified model. However, simply assuming the maximum number of states and subsequently disregarding the phantom states by default is not advised. First, reconstruction of a higher number of states requires experimental RDC data from more alignment media. While it is always encouraged to acquire as much RDCs as possible, this may not always be feasible. Second, collectively a few phantom states (all individually negligible) may conceivably represent one real state. In general, it is therefore recommended that phantom states be eliminated one at a time followed by reevaluation.

## Chapter 3: Summarization of Major Contributions for the Creation and Improvements of Methods to Calculate Protein Structure and Dynamics

### 3.1 Improving the Usability of REDCRAFT

The first objective of this part of my research was improve the existing software REDCRAFT to enhance its abilities to calculate protein structure. A full description of the work completed can be found in section 2.1. The following is a summary of the major achievements. First and foremost, the interface to the software has been updated to include a modern GUI and the code refactored to allow the users to import a larger variety of experimental data. These updates were aimed at increasing ease of use and spur more widespread adoption of the software. Runtime analysis of the software before and after refactoring was completed and concluded there to be no significant reduction in the speed or accuracy of calculation. Secondly several of its computational features were enhanced. The decimation routine, responsible for reducing the solution space of the algorithm, was updated to dynamically scale its sampling of clusters instead of statically selecting samples based on user input (pseudocode in Appendix A). This update has now allowed for more complicated datasets to be utilized in the software. For example, in the previous version of REDCRAFT, structure calculation using highly erroneous data would quickly exhaust the RAM of even some of the most well-equipped computers (64 GB). With the new version of decimation, structures are able to be characterized using these datasets in a matter of hours. Lastly, Stage-I of the REDCRAFT algorithm was improved by adding the ability to

incorporate of dihedral angle restraints mined from a newly created database (PDBMine discussed in section 2.2). In previous versions of REDCRAFT Stage-I used the entire allowed Ramachandran space. Even though this removed some of the theoretical solution space (-180 to 180) it still left a significant area for the algorithm to search through. Using these enhanced Ramachandran spaces mined from PDBMine, protein structures were computed from sets of experimental data (RDCs) that were previously thought impossible. For example, it has long been accepted in the field of NMR spectroscopy that structure calculation from RDCs alone required using data from at least two alignment media to resolve inherent rotational degeneracies. Using REDCRAFT alongside mined dihedral angle restraints, a high-resolution structure (within 2.4 angstroms) was calculated using just one set of N-H RDCs. Further work is needed in this study to confirm these results in a larger variety of proteins but if irrefutably confirmed, this could be a major turning point for the field of NMR and spur a more cost-effective pipeline of structure calculation.

In addition to the published works contained in Chapter 2 of this document, I have made other noteworthy contributions to the field of protein structure calculation. Using the enhanced version of REDCRAFT, I have been able to characterize a novel protein, Pf2048.1 (PDB IDs: 6E4J, 6NS8). At the time of its calculation, it had less than 11% sequence similarity with any other known protein. It's inclusion into the Protein Databank means a reduction in the existing gaps in the human proteome.

### 3.2 Data Driven Dihedral Angle Restraints

The second objective of this part of my research was to create a minable version of the Protein Databank. Whereas the full description of results can be found in section 2.2, the following is a summary of the major achievements made to this end. The protein databank

(PDB) was dissected and used to create the PDBMine database. The new database occupies over 300 GB of space and contains atomic and dihedral level information for over 400,000 proteins. Whereas the original database platform was MySQL (as reported in the published work), it has since been updated to a binary file in which queries are made by using a jump table to index into various locations. The wrapper to the database is written in python and allows for easy and flexible mining of the data. The results from various queries of this database have been shown to greatly enhance protein structure prediction methods. This has been reported in several papers (sections 2.1 and 2.2) as well as in publications currently in preparation. Currently, rudimentary website is hosted at the following URL: <https://www.ifestos.cse.sc.edu/PDBMine> and allows for mining of only the dihedral angles. The website implements a RESTful interface for handling queries from users and returns queries of ~100 amino acid sequences at a window size of 7 in approximately 5 minutes. Future development of the user interface is underway to enable the full set of features available the python command line interface.

### 3.3 Creation of RDC-based Model of Protein Dynamics

The third objective of this part of my research was to formulate an analytical method of describing atomic level models of protein dynamics. The following is a summary (full work in section 2.3) of the major achievements made to this end. An objective function was formulated and utilized to reconstruct 2-state and 3-state discrete dynamic systems. This serves as the first ever atomic level modelling of discrete protein dynamics. The objective function takes advantage of the averaging effect seen in RDCs collected from dynamical protein in NMR experiments. In these systems, there are two clear domains of the protein moving with respect to one another. In our experiments, one

domain was arbitrarily chosen as the “static” domain and the other was designated the “dynamic” domain. An order tensor was then calculated for each domain and the objective function was then used to relate the two domains together using a series of Euler rotations and occupancy rates. This model showed success on models with motions as small as 10 degrees and rates of occupancy as low as 20 percent. Although the work presented here was all completed on synthetic data, many real-world systems exhibit similar motion to the motion that was simulated and are then therefore likely to also be able to be characterized using this model.

### 3.4 Suggestions for Future Work

My suggestions for future work in this line of research would focus on additions to the theoretical models, enhanced testing on real world protein systems and expansion of the PDBMine interface. The model created to characterize discrete dynamics can be improved in several ways. The first of which is to investigate and resolve seemingly inherent degeneracies of the objective function. During the course of investigation, a few anomalies were encountered wherein a certain combination of order tensors, degrees of motion and occupancy rates caused the function to get stuck in local minima. The second suggestion would be to apply the model to real world models of dynamical protein systems. There are several models that exist however RDC data is not currently available to study them using this model. Lastly, the interface of PDBMine should be extended to provide a

## Chapter 4: Introduction to Smoking Detection

### 4.1 Smoking Behavior and How it is Traditionally Studied

Any study of a specific human behavior will rely, in some form, on self-reporting. In the realm of smoking research, it is no different. Many behavioral studies centered around smoking still rely on some form of survey that the participant needs to complete on a given day describing parameters such as: number of cigarettes smoked, time of the day when most smoking occurred. This traditionally was performed utilizing pen-and-paper logs that the participants would return to the researchers at the end of the trial period. This method led to a high degree of data loss from human error (either forgetting to log or logging an event incorrectly). In recent years, mobile technology has been utilized to facilitate a more conducive self-reporting experience and minimize retrospective recall[170]. A Google form replaced the pen-and-paper logging method, allowing researchers to view data on more regular intervals to ensure that participants were adhering to protocol. Mobile technology also has the capability of providing detailed contextual information that can be used to better inform prediction models. For instance, smart watch devices have both an accelerometer as well as GPS. Using these sensors an intelligent system could be devised to listen for key events, such as consumption of fast food, that may trigger a more destructive behavior such as smoking. In addition, location and time information can be used in combination to create a probabilistic model of behavior and therefore inform the system when a person is most likely to partake in a destructive behavior ahead of time such



that “in time” intervention can be administered. Although the use of mobile devices and web-based reporting techniques improve the quality of participant-reported data along with providing more contextual information, on their own, they fail to mitigate the burden of self-reporting and the errors associated with subjective reports[166, 171, 172].

One highly effective way to mitigate the burden of self-reporting is to collect behavioral data in the laboratory setting. Clinical Research Support System for Laboratories (CReSS) is widely used for studying smoking in a laboratory setting. The device collects several measures of smoking behavior including puff duration, distance between puffs (Inter Puff Interval), time to first puff, and puff volume. However, the CReSS device is an expensive (~\$5,500) machine that interferes with the natural smoking experience. When using the CReSS, the cigarette is fitted into a port on one side of the device and smokers inhale through a plastic mouthpiece. Participants often complain about how smoking through the CReSS feels strange and does not provide the same sensations as experienced through naturalistic smoking. In addition, the study of smoking with CReSS forces the laboratory confinement of the experiment. The study of smoking in laboratory settings provides skewed data, since the participants may prefer to complete the experiments quickly, in uncomfortable circumstances and times of the day. Use of mobile devices, specifically wearable devices, hold the potential to eliminate these limitations.

## 4.2 Previous Work

Previous works have shown the possibility of detecting smoking gestures using in-house designed wearable devices[173, 174]. These techniques have shown great promise with both high accuracy and low false positive rates. However, they require the use of devices not commonly found in a typical household, such as multiple 9-axis inertial

measurement units (IMU's), respiration bands that must be worn across the chest, and two-lead electrocardiograph sensors worn under clothes. The use of these uncommon, cumbersome, and relatively expensive devices severely limits mass deployment for daily observations or use by the research community. Recently, smartwatches have been explored as a platform for observing people's daily activities. Specific to the smoking activity, StopWatch[175] and SmokeBeat[176] independently confirmed the possibility of utilizing smartwatches in the study of smoking. StopWatch reported a precision rate of 86% and a recall of 71% in detection of smoking in natural settings where as SmokeBeat reported performance of 82% (due to its commercialization publications related to SmokeBeat are limited in detail). The StopWatch algorithm employs a decision tree model and requires preprocessing of the data for feature engineering. While feature engineering and data transformation may provide certain advantages, they impose additional compute cycles that will be taxing on the limited battery resources of smartwatches.

### 4.3 Research Objectives

The work presented here features the usage of only raw accelerometer data and a combination of shallow and deep neural networks to detect and characterize smoking in a variety of ways. Of primary interest was detection of smoking in natural settings for real time intervention and the development of a system to characterization of smoking behavior to replace laboratory devices and enable the study of smoking in natural settings. Details of each objective are outlined in the following sections.

#### 4.3.1 - Puff Level Binary Detection of Smoking

A puff is defined by the motion of the hand from a resting position to the lip and then back to the resting position. To detect these gestures, a variety of ANNs trained using

accelerometer data collected from various smoking and non-smoking events were designed, tested and deployed. The main objective for this level of detection was to perform binary classification (smoking or non-smoking) in order to facilitate real time intervention measures. This binary classification (“smoking” vs. “non-smoking”) was completed using a traditional ANN. The network was trained with smoking data collected in a laboratory setting. The “true positive” input to the network were windows of three seconds of accelerometer data that contained a full puff gesture. The “true negative” input to the network was extracted from a variety of non-smoking activities such as eating, drinking and exercising. The network was then tested rigorously using laboratory collected data as well as data from real smokers in natural settings. In this preliminary, proof of concept work, data from only one real smoker was used for validation. This data was then annotated by a highly trained researcher and then tested using the network previously trained using the lab setting data. The results for this work can be found in the section titled: “5.2 Recognition of Smoking Gesture Using Smart Watch Technology.” These results were originally published in 2016 at the International HIMS conference.

#### *4.3.2 - Session Level Detection of Smoking*

The main objective in this area of research was to identify smoking at the session, or full cigarette, level. The implementation of this objective partially relies on the success of the first objective. In some studies, session detection is much more valuable than detection at the puff level. For instance, many cessation studies simply focus on the number of cigarettes smoked in a given day. For this type of study, the puff level detection does not have to be as precise. Session level detection is also important in the context of other types of tobacco related research such as the effectiveness of warning labels. In these instances,

the researchers are interested in detecting a full session of smoking (completion of one cigarette) before initializing a short survey to the smoker to inquire about the label that was included in the cigarette package.

The main first challenge of this objective was that a smoking session was not analytically defined. Therefore, the first task was to define a session. Utilizing both subject matter experts and empirical data collected from real smokers, a model of a smoking session was developed. This model is defined as a collection of three or more puffs that occur with an inter-puff interval of less than four minutes in an overall time frame of less than 8 minutes. In addition, two sessions must be separated from one another by five or more minutes. Session level detection was accomplished by utilizing the results from the puff level detection in combination with a rule-based AI that will confirm the adherence of the ANN output to the model defined.

Evaluation of session level detection was performed by recruiting ten heavy smokers and continuously collecting smartwatch accelerometer data from them in real-time. The participants also utilized a Google Form as a mechanism of marking the start and end times of each smoking session throughout the day. Using the data, the session model and the output of the ANN from the first objective, the accuracy of session level detection was computed. The results from this work can be found in the section titled: “5.3 Detecting Smoking Events Using Accelerometer Data Collected Via Smartwatch Technology: A Feasibility Study” which was originally published in JMIR uHealth mHealth in 2017.

#### *4.3.3 – Automated Characterization of Smoking Topography*

The next research objective was the characterization of smoking topographies. This task required an approach different to that of the previously described binary puff

classification due to its added complexity. For example, to accurately calculate the puff duration, at least two points of each puff will need to be detected: the hand-to-lip gesture and the hand-off-lip gesture. The duration can then be calculated as the time between these two measures. To accomplish this the data was reformulated into classification of “mini” gestures. After successful reannotation of the data, a conventional neural network was employed that had an output size of four (“hand-to-lip”, “hand-on-lip”, “hand-off-lip” and “non-smoking”). The output of this network was then fed to a rule-based AI that modelled the smoking behavior as a state transition model. As a prerequisite to deploying a full neural network on the problem, proof of concept study was performed. The results of which can be found in the section titled “5.4 Clinical Quantitation of Smoking Topography using Smartwatch Technology”. These results are currently under review at JMIR. In this work, smoking data collected from a smartwatch was compared to that of data collected from the CReSS smoking device. Several measures such as duration and inner puff interval were compared using  $R^2$  values.

#### *4.3.4 - Resolving Ambiguities in Accelerometer Data Based on the Position of Smartwatch on Wrist*

The final research objective was to develop a translational mechanism that would account for the rotational discrepancies in accelerometer data based on the position of a smartwatch on the wrist. This is a very important step in processing the data for use in a ANN. There are eight distinct configurations of watch placement on the wrist which lead to variations in the accelerometer data. Instead of insisting on a particular configuration to be used across all users, a transformation technique will be investigated to rotate all data into a common frame to be used within the network. The transformation technique can be

evaluated in a straightforward fashion. Data will be collected in each possible configuration and then transformed. The resulting data will then be aligned and compared to data collected in a given common frame. The common frame that will be assumed for this evaluation will be right side up in a pronated position on the left wrist. The results of this study are shown in the section entitled: “5.5 Resolving Ambiguities In Accelerometer Data Due To Location Of Sensor On Wrist In Application to Detection of Smoking Gesture.”

## Chapter 5: Contributions to the Study of Automated Smoking Detection<sup>2</sup>

---

<sup>2</sup> Publications in this Chapter:

Section 5.1: Casey A. Cole, Bethany Janos, Dien Anshari, James F. Thrasher, Scott Strayer, Homayoun Valafar. 2016. Proceedings of the International Conference on Health Informatics and Medical Systems (HIMS). Reprinted here with permission from the publisher.

Section 5.2: Casey A. Cole, Dien Anshari, Victoria Lambert, James F. Thrasher and Homayoun Valafar. 2017. JMIR Mhealth Uhealth 2017;5(12):e189. Reprinted here with permission from publisher.

Section 5.3: Casey A. Cole, Shannon Powers, Rachel L. Tomko, Brett Froeliger, Homayoun Valafar. Submitted to JMIR.

Section 5.4: Casey A. Cole, James F. Thrasher, Scott Strayer, Homayoun Valafar. 2017. Contributed paper at 2017 IEEE International Conference on Biomedical and Health Informatics. Reprinted here with permission from the publisher. See IEEE copyright: <https://www.ieee.org/publications/rights/copyright-policy.html> © 2017 IEEE

## 5.1 Recognition of Smoking Gesture Using Smart Watch Technology

### 5.1.1 Abstract

Diseases resulting from prolonged smoking are the most common preventable causes of death in the world today. In this report we investigate the success of utilizing accelerometer sensors in smart watches to identify smoking gestures. Early identification of smoking gestures can help to initiate the appropriate intervention method and prevent relapses in smoking. Our experiments indicate 85%-95% success rates in identification of smoking gesture among other similar gestures using Artificial Neural Networks (ANNs). Our investigations concluded that information obtained from the x-dimension of accelerometers is the best means of identifying the smoking gesture, while y and z dimensions are helpful in eliminating other gestures such as: eating, drinking, and scratch of nose. We utilized sensor data from the Apple Watch during the training of the ANN. Using sensor data from another participant collected on Pebble Steel, we obtained a smoking identification accuracy of greater than 90% when using an ANN trained on data previously collected from the Apple Watch. Finally, we have demonstrated the possibility of using smart watches to perform continuous monitoring of daily activities.

### 5.1.2 Introduction

In the past decade, measures have been taken to warn the population about the dangers of smoking. While the smoking rate has decreased significantly since then, smoking remains the leading preventable cause of death throughout the world. Additionally, youth tobacco use has increased as the popularity of products such as e-cigarettes and hookah has risen[177]. In America, 53.4% of college students have smoked at least one cigarette and 38.1% reported smoking in the past year[178]. Even though the



hazards of smoking are generally accepted, there remains many smokers who struggle to quit. Those who try to quit are typically middle aged and beginning to feel the adverse effects of smoking. Yet, on average, smokers relapse four times before successfully quitting[179]. Many smokers do not realize that it is normal to require multiple attempts to quit smoking and therefore need recurring intervention and support to aid them. Constant support from an individual's community is shown to increase the likelihood of quitting[178]. The existence of an application (housed on a smart phone or watch) that would provide this constant support could greatly increase a person's fortitude to abstain from smoking.

The first step in making such an application is the ability to detect when a person is smoking so that the appropriate intervention can be initiated. Previous works have shown the possibility of detecting smoking gestures using in-house designed wearable devices[173, 174]. These techniques have shown great promise with both high accuracy (95.7-96.9%) and low false positive rates (<1.5%). However, they require the use of devices not commonly found in a typical household such as multiple 9-axis inertial measurement units (IMU's), respiration bands that must be worn across the chest and two-lead electrocardiograph worn under the clothes. The use of these uncommonly and relatively expensive devices severely limits mass deployment for daily use.

Smart watches are becoming increasingly prevalent in common households. According to Apple's website, over 5 million Apple Watches were sold in 2015 alone and projections into 2016 show promising growth. Other smart watch companies like Asus and Pebble have seen similar growth patterns from as well. By contrast to the previous methods,

the method explored in this study relies solely on the use of a smart watch's built-in accelerometer, effectively eliminating the need to use more uncommon detection devices. In addition, the pairing of a smartwatch with a smart mobile device enables immediate alerting, engagement and recruitment of social support groups to prevent or alter one's smoking behavior. As the first step in this process we have examined the feasibility and complexity of detection of smoking gesture using smart wearable devices. Our investigation has included minimum data requirement and an exploration of most informative dimension of accelerometer sensors. Prior knowledge of the problem complexity will allow for a smoother transition into actual deployment of our detection mechanism on smart watches in the future.

### *5.1.3 Background and Method*

The overall view of our study consisted of three major stages: data collection, training of multiple artificial neural networks for pattern recognition, and evaluation. The following sections provide a more detailed outline for each of these stages.

#### *5.1.3.1 Data Collection*

Data in this study were acquired by a non-smoking participant utilizing an Apple Watch (version 2.1). Using the application PowerSense (available in App Store for iOS) a number of individual smoking gestures (also referred as a puff) and continuous smoking sessions (a session that consists of multiple puffs) were recorded and analyzed. All samples were measured at a 50Hz sampling rate. Due to minor fluctuations in the duration of each gesture, the number of data points varied for each gesture. Each isolated puff pattern was represented by 200 interpolated data point in order to create a uniform size input set for the pattern recognition stage. The resulting smoking gestures are shown below in Figure 5.1.

The three differently colored line clusters represent each of the three dimensions of the accelerometer (X in blue, Y in red and Z in green). Each cluster is an overlay of all 20 individual smoking sessions used in the training of the neural networks. Based on visual inspection, it is clear that the smoking pattern is very well conserved across each of the samples.

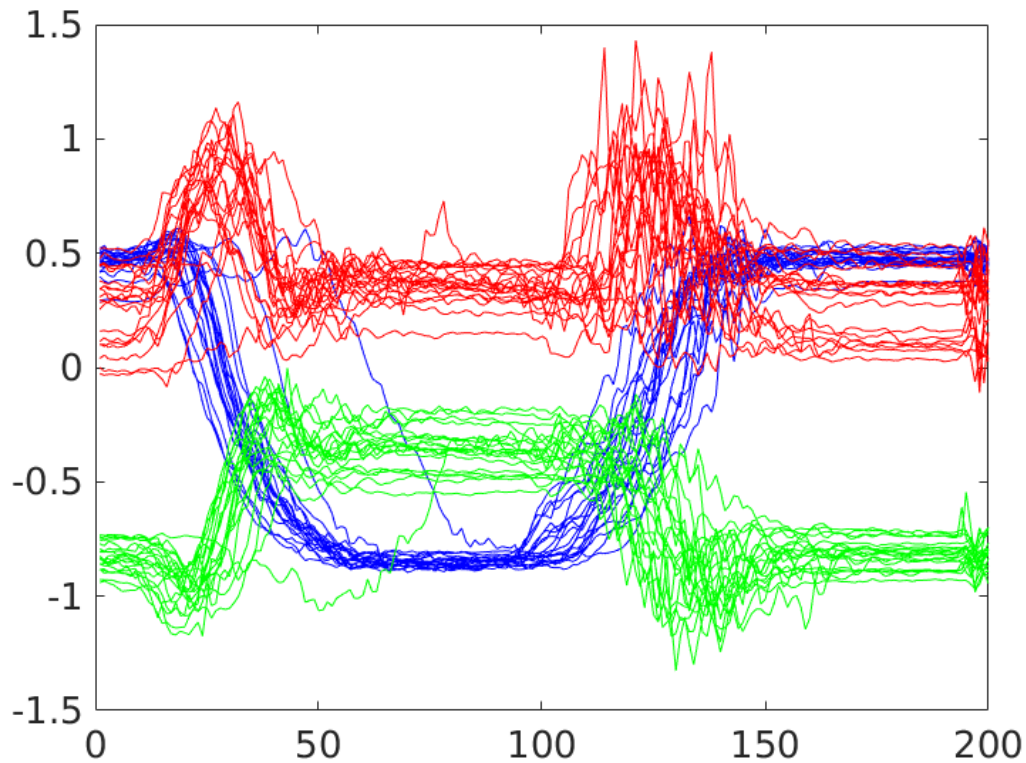


Figure 5.1. Overlay of all single smoking gestures with the X dimension in blue, Y in red and Z in green.

In addition to smoking sessions, several non-smoking gestures were also recorded. These gestures (seen in Figure 5.2) included drinking, scratching one's nose, yawning, coughing, brushing hair behind one's ear and rubbing one's stomach. The selection of these patterns was based on activities that may be similar to smoking gesture, or ones that may

be present during most common smoking sessions. These gestures were including in both the training test and testing set.

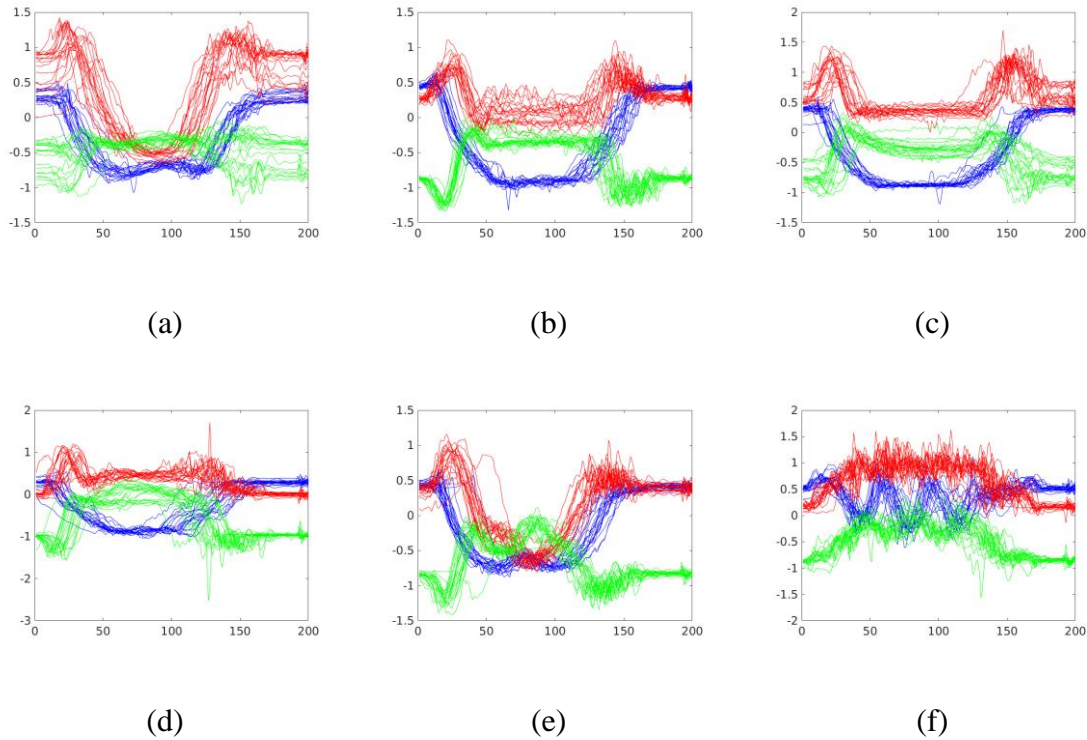


Figure 5.2. Overlay of all patterns collected for the following non-smoking gestures: (a) drinking, (b) scratching one's nose, (c) yawning, (d) coughing, (e) brushing hair behind one's ear, and (f) rubbing one's stomach.

In some of these gestures, such as scratching one's nose and yawning in Figure 5.2(b) and Figure 5.2(c) respectively, the movement of the hand and arm clearly resemble an individual smoking gesture (seen in Figure 5.1). Inclusion of these gestures into the data set will allow for studying how well the proposed method can distinguish smoking gestures from other very similar gestures.

In addition to individual gestures, longer continuous sessions were recorded. The continuous smoking sessions consisted of approximately 7 to 10 gestures per session. The non-smoking sessions included three common activities: eating, drinking and putting on

chapstick/lipstick. Each session was divided into 200 time step segments using a rolling window approach. As seen in Figure 5.3, a continuous smoking session is no more than a combination of individual smoking gestures seen in Figure 5.1.

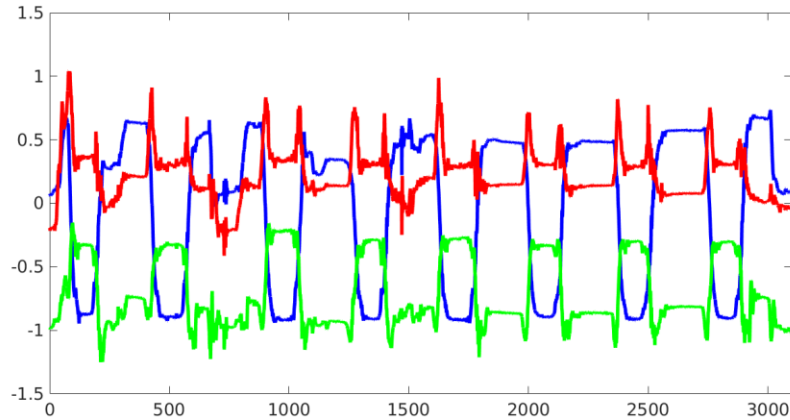


Figure 5.3. Example of continuous smoking session.

Table 5.1 summarizes the total number of smoking and non-smoking gestures in each of the data sets and breaks down the exact amount used in each phase of the investigation.

Table 5.1. Summary of data utilized in analyses.

	Non-smoking	Smoking
Training Set	120	20
Testing Set	30	10
Extended Sessions	5	5

In order to test the applicability of the presented method across other wearable platforms, as well as other participants, one smoking session was acquired by a different participant (than the original set of data used for training) on the Pebble Time Steel used on Android platform. The Pebble Time Steel was selected as a second test watch due to its

long lasting battery life, durability, and reasonable price. A continuous smoking session was recorded using the AccelTool (<http://mgabor.hu/accel/>) App at a sampling rate of 50 Hz.

### 5.1.3.2 Pattern Recognition Via Artificial Neural Networks

The neural network toolbox in Matlab (version R2016a) was utilized during this phase of our study. Levenberg-Marquardt backpropagation[155, 180, 181] was selected as the training algorithm in all sessions. For each training session, the data were randomly partitioned into 3 sets: 70% in the training set, 15% in the validation set and 15% in the testing set. The networks were then trained, validated and then rigorously tested for accuracy. The procedures for the training and validation/testing phases are outlined below.

#### 5.1.3.2.1 *Training*

The interpolated raw data consisted of information from three dimensions (X, Y and Z). In order to fully identify and evaluate useful information in the data, all three dimensions were utilized both individually and in combination with each other. In total, five ANNs were created for use in this study—one for each of the three dimensions (X, Y and Z), one for the combination of all three dimensions (referred to as XYZ) and one for the average of the three dimensional data (referred to as AVG). The number of inputs for the X, Y, Z and AVG data sets was 200, while the input size for the XYZ data set was 600. In each of the neural networks the hidden layer consisted of 10 hidden neurons. A single output neuron was used, with zero denoting a non-smoking gesture and one signifying a smoking gesture.

#### 5.1.3.2.2 Validation/Testing

Validation of the appropriate level of training was accomplished using 15% of the excluded training dataset. The networks were further subjected to testing using several different data sets. The first of which was the remaining 15% of the data excluded from the original training set. Next, a new set of individual gestures (not included in the original training set) was presented to the networks. To test the method on more realistic cases, continuous smoking and non-smoking sessions were also presented to the networks. Lastly, a continuous smoking session from a different smart watch was tested on each of the ANNs. The results of each test set are reported in Section 5.1.3.

#### 5.1.3.3 Evaluation

To measure the success of the proposed method specificity, sensitivity and total accuracy of each trial were observed. Specificity describes the rate at which the method is able to correctly classify a non-smoking event. Specificity is calculated by use of Eq (5.1.1), where  $TN$  and  $FP$  denote the number of true negatives and false positives, respectively.

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (5.1.1)$$

Sensitivity refers to the rate at which the method correctly identifies a smoking event and can be calculated using Eq (5.1.2), where  $TP$  represents the number of true positives and  $FN$  denotes the number of false negatives.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (5.1.2)$$

Total Accuracy is then a measure of how often the method correctly classifies both smoking and non-smoking gestures and is calculated by Eq (5.1.3).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5.1.3)$$

In this work a cutoff threshold of 0.8 (or 80%) was used above which was considered a successful detection.

#### *5.1.4 Results and Discussion*

In the following sections the results of testing the neural networks are reported. In each section the data set that performed the best and worst are discussed.

##### *5.1.4.1 Accuracy on Training Set*

For each of the data sets the corresponding neural networks was independently trained 10 times and the network with the highest accuracy was chosen to be used for future test sets. The accuracies, specificities and sensitivities described in Figure 5.4 represent the performance of the final trained neural networks on their respective training sets.



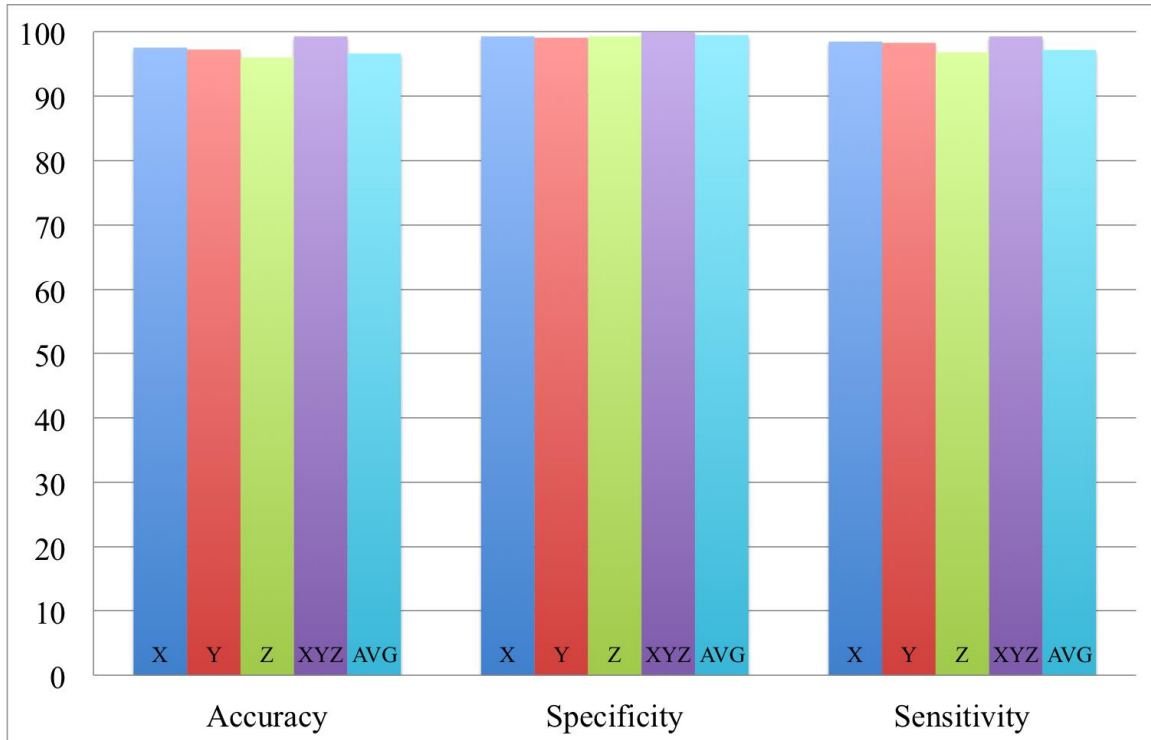


Figure 5.4. Accuracy, specificity and selectivity of the neural networks during training. The bars are individually labeled based on their respective training sets.

All of the neural networks exhibited high accuracy and specificity (> 90%) with respect to the training set of data. In each case the network trained with the XYZ data performed the best and the network trained with just the Z data performed the worst. A visual comparison of the individual smoking gestures and the non-smoking gestures clearly explains Z's poorer performance in correctly classifying smoking gestures. The Z dimension (in green Figures 5.1 and 5.2) exhibits very similar patterns across both smoking and non-smoking gestures. However, it is worth noting that the Z dimension still obtained a high specificity (nearly 100%) which means that it still may carry some complementary information, especially when identifying a non-smoking gesture.

#### 5.1.4.2 Individual Gesture Detection

In this experiment, a new set of individual gestures (smoking and non-smoking) were presented to the previously trained neural networks. The accuracy, specificity and sensitivity are reported in Figure 5.5. Accuracies were measured by forward propagating each of the new samples in the corresponding neural network and then recording the number of correct and incorrect predictions. A threshold of 0.5 was used in interpretation of the neural network output, that is, any output larger than 0.5 was considered as smoking and any output lower than 0.5 was considered as non-smoking.

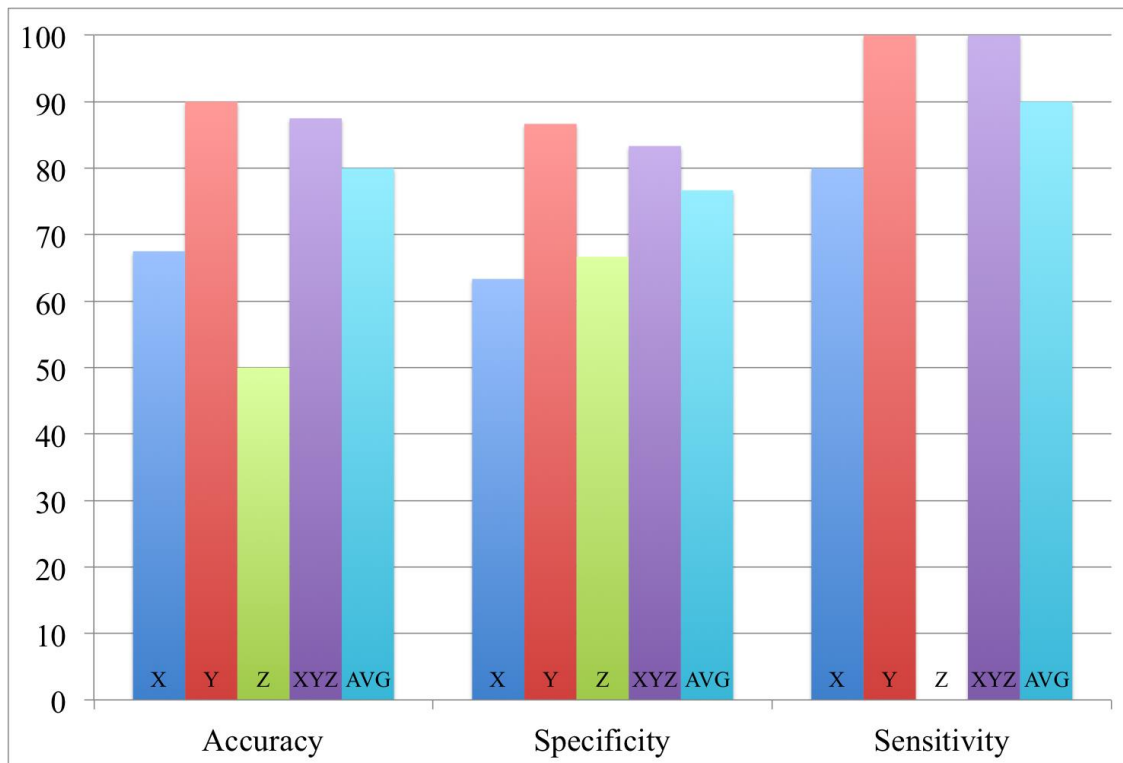


Figure 5.5. Accuracies, specificity and sensitivity in the individual gesture detection trials.

As shown in Figure 5.5, the Y dimension performed the best with not only the highest accuracy, but also the highest specificity and a 100% sensitivity. The XYZ and

AVG data sets also performed well especially in their ability to identify smoking gestures. Consistent with the previous section, the Z dimension performed the worst with both a low accuracy and specificity as well as a 0% sensitivity rate indicating utilization of the Z dimension results in identification of 0/10 smoking gestures.

To better understand the nature of false-positive classifications, contribution of each individual gesture was recorded and results are shown in Figure 5.6. Based on the results shown in this figure, the non-smoking pattern that caused the most false positives was coughing followed by scratching of nose and yawning. These results were expected due to the high degree of visual similarity of these gestures to an individual smoking gesture.

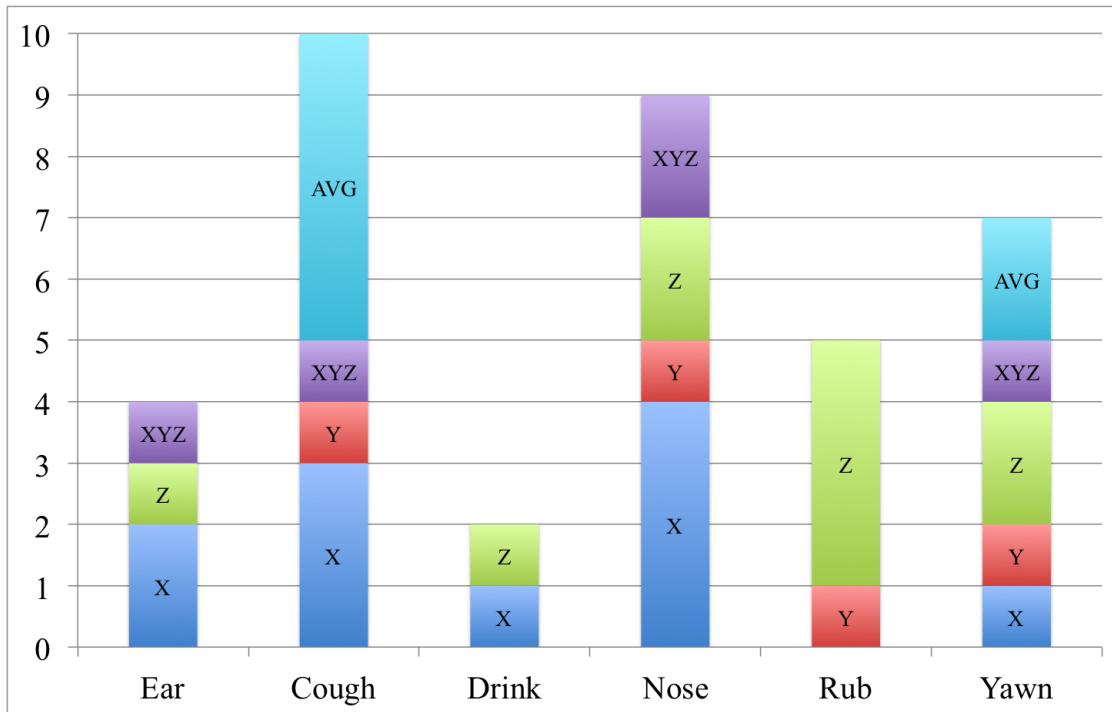


Figure 5.6. Total number of false positives created by each non-smoking gesture. Each segment is labeled based on the dimension of the accelerometer data being used.

#### 5.1.4.3 Continuous Gesture Detection

In this section the results for detection on continuous monitoring of gestures are reported. To accomplish this objective the neural networks trained on static gestures were utilized. Using a running window of size 200 (without any interpolation) the continuous gestures were parsed into data sets for input into the networks. The classification result for each running window (0 denoting non-smoking and 1 denoting smoking) is plotted at the beginning of each running window. Figure 5.7 illustrates an example output (in purple before converting to a binary representation) of the neural network trained on X. Visual inspection of this figure clearly confirms the correlation between spikes in detection pattern over the regions where an apparent smoking gesture. However, there is not a trivial way to quantify the network's success because it is not immediately clear where should constitute the start and end of a gesture. Therefore, in order to be sure to encompass the entire gesture, generous ranges were handpicked to describe each smoking gesture. As in the previous section, a cutoff of 0.5 was chosen where any output greater than 0.5 was considered smoking and anything below was counted as non-smoking. Specificity was measured by considering all other sections of the continuous gesture not within the smoking ranges.

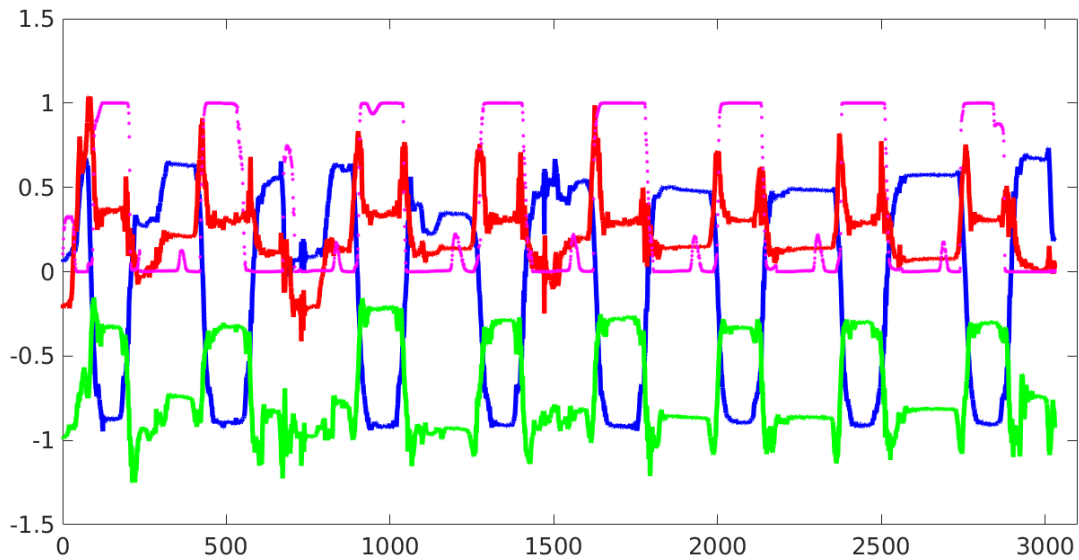


Figure 5.7. Example of continuous smoking session superimposed with the output of the neural network trained on the X dimension.

The averaged results across all five continuous sessions are presented in Figure 5.8 along with error bars representing the minimum and maximum of each averaged result.

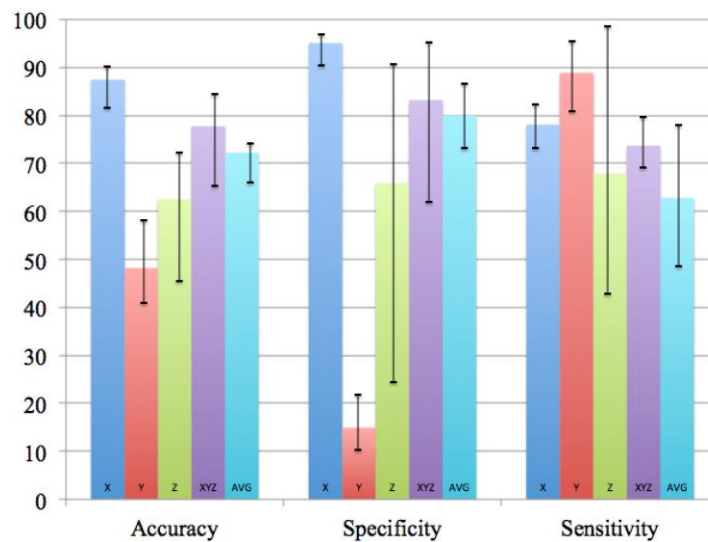


Figure 5.8. Averaged accuracies, specificities and sensitivities across the five continuous smoking gesture detection experiments with error bars representing the respective min and max of each value.

In the continuous smoking sessions, the X dimension performed better than all other dimensions. The Y dimension had a better sensitivity score but this can be disregarded due to its low specificity score. A high sensitivity coupled with a low specificity score denotes that the Y dimension classifies practically everything as smoking and therefore its high sensitivity should be ignored. In this sense, Y performed the worst overall.

In the non-smoking sessions selectivity becomes inapplicable and specificity is equivalent to total accuracy. therefore, only accuracies these sessions are presented in Figure 5.9.

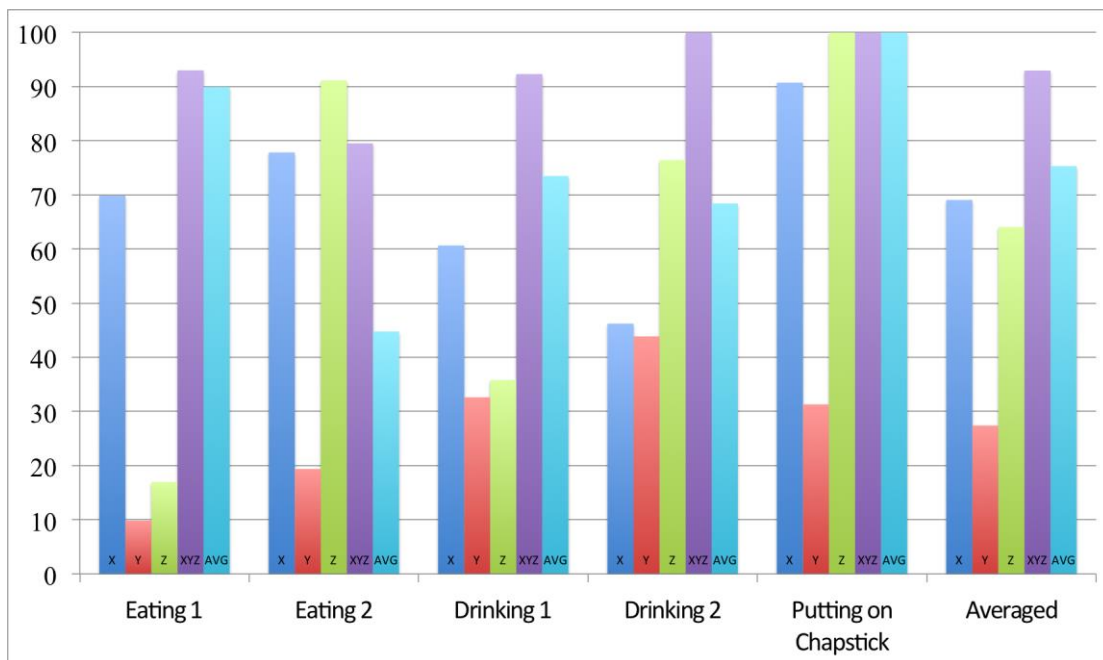


Figure 5.9. Accuracies for five continuous non-smoking session trials.

Despite its superior performance in classification with continuous smoking gestures, the X dimension performed with an average accuracy of under 70% in the non-smoking continuous sessions. It seems that in the presence of more complex continuous

gestures (like eating and drinking) the XYZ data set seems to contain the most useful information for correct classification. Eating sessions seemed to pose the most difficulty for XYZ. This could signify that eating is one of the closest gestures to smoking and can therefore lead to confusion in the network. In accordance with previous results, the Y dimension performed the worst across all the non-smoking sessions.

#### 5.1.4.4 Exploration of dependency on the wearable device

As described in the Section 5.1.2.1, a continuous smoking gesture was recorded using a Pebble smart watch. Results were collected using the pre-existing neural networks that was trained on data from the Apple Watch from a different participant. Figure 5.10 shows the smoking session recorded using the Pebble watch, which exhibit significant similarity to patterns shown in Figure 5.1. The outputs of the neural network using the X data set are shown in purple. Again, there is good correlation between the smoking gestures and the spikes in the output of the neural network. Figure 5.11 shows the resulting accuracies, specificities and sensitivities for this smoking session.

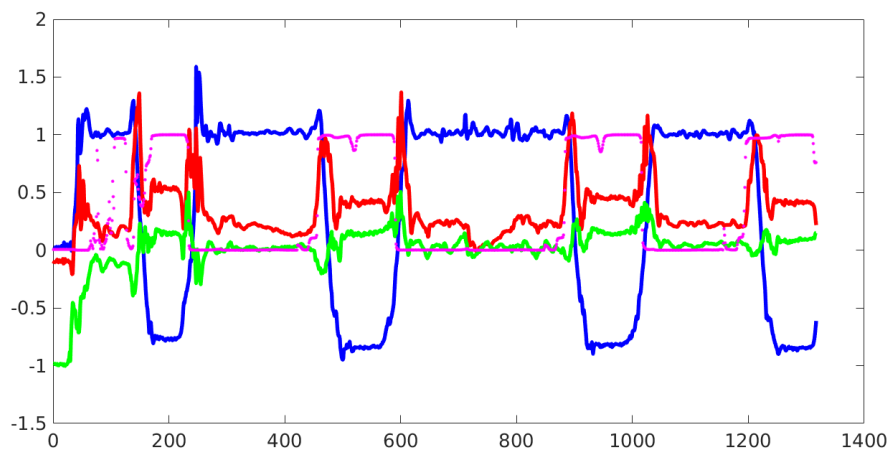


Figure 5.10. Output of the neural network for the X dimension superimposed to the original smoking session.

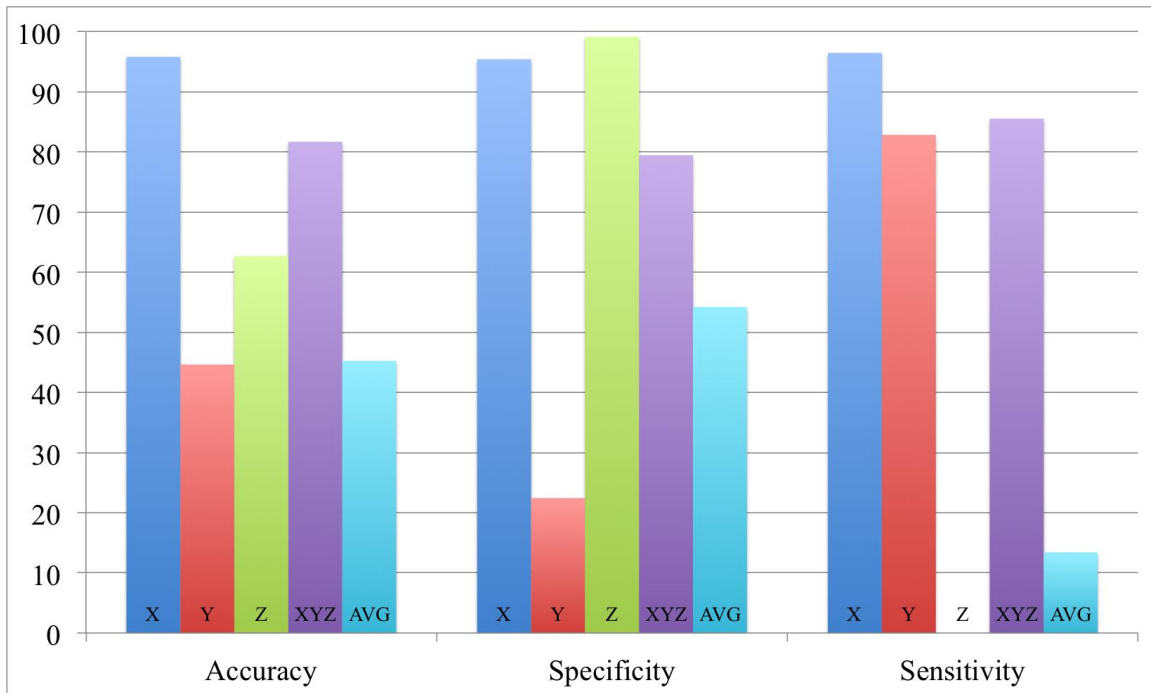


Figure 5.11. Results for the Pebble smart watch.

As in the previous cases of continuous smoking sessions, the X dimension performed the best in classifying the gestures. In the case of the Pebble watch, the Y and Z dimensions did equally poorly. The Y dimension classified almost everything as smoking and the Z dimension classified nearly everything as non-smoking.

### 5.1.5 Conclusion

The general summary of our work supports the feasibility in detection of smoking gestures using typical sensors available in smart watches. Based on our experiments, pattern recognition via Artificial Neural Networks applied to the sensor data obtained from smart watches can produce performances very comparable to previously reported work. However, the use of a smart watch is far more pragmatic in general population studies over the other existing technologies. Results shown in section 5.1.3.4 suggest the possibility of delivering an application capable of detecting smoking gesture across the general



population of smokers. This universal Artificial Neural Network eliminates the need to customize a training session per user.

Our exploration of efficacy of individual sensor data in detection of gesture has produced unexpected results. The neural network trained with data from just the X dimension performed the best in the presence of continuous smoking gestures but when faced with more complex non-smoking motions, it fails and a more complete set of data is needed to distinguish smoking gestures. Across all the testing sets the neural network trained with data from all three dimensions (XYZ) did consistently well. However, the XYZ data set requires 600 inputs to the network whereas the X data set only requires 200. This is a significant reduction in data requirement which directly impacts the computational time of the method. For this reason, the viability of both data sets will continue to be investigated.

Additional investigations are required before general deployment of such approaches. Continuous monitoring of data may be outside of power limitations of such devices and may act as a technological barrier. Our future investigations will include optimization of sampling rate, minimization of Bluetooth communication between the smart watch and the companion phone, and better assessment of the universality of the trained ANN.

## 5.2 Detecting Smoking Events Using Accelerometer Data Collected Via Smartwatch Technology: A Feasibility Study

### 5.2.1 Abstract

Smoking is the leading cause of preventable death in the world today. Ecological research on smoking in context currently relies on self-reported smoking behavior. Emerging smartwatch technology may more objectively measure smoking behavior by automatically detecting smoking sessions, using robust machine learning models. This study aimed to examine the feasibility of detecting smoking behavior using smartwatches. The second aim of this study was to compare the success of observing smoking behavior with smartwatches to that of the conventional self-reporting. A convenience sample of smokers was recruited for this study. Participants (N=10) recorded twelve hours of accelerometer data using a smartphone and smartwatch. During these twelve hours, they engaged in various daily activities including smoking, for which they logged the beginning and end of each smoking session. Raw data were classified as either smoking or non-smoking using a machine learning model for pattern recognition. The accuracy of the model was evaluated by comparing the output with a detailed description of a modeled smoking session. In total, 120 hours of data were collected from participants and analyzed. The accuracy of self-reported smoking was ~78% (96/123). Our model was successful in detecting 100 out of 123 (~81%) smoking sessions recorded by participants. After eliminating sessions from the participants that did not adhere to study protocols, the true positive detection rate of the smartwatch based-detection increased to more than 90%. During the 120 hours of combined observation time, only 22 false positive smoking sessions were detected resulting in a 2.8% false positive rate. Smartwatch technology can

provide an accurate, non-intrusive means of monitoring smoking behavior in natural contexts. The use of machine learning algorithms for passively detecting smoking sessions may enrich ecological momentary assessment protocols and cessation intervention studies that often rely on self-reported behaviors and may not allow for targeted data collection and communications around smoking events.

### *5.2.2 Introduction*

Despite rapid adoption of many tobacco control policies around the world, cigarette smoking remains the greatest preventable cause of death[163]. Ecological momentary assessment studies are increasingly popular for understanding smoking behavior in context[182-184]. Studies in this area have traditionally relied on participants to self-report smoking behaviors in real time, which can be particularly burdensome for heavier smokers and result in missing or biased information if participants are not forthcoming about or forget smoking events[185]. In this study, a smoking event can be either an individual puff or an entire session, where a session is defined to be the time in which it takes to smoke a single cigarette. Emerging technologies that allow for passive detection of stereotyped behaviors like smoking may be able to decrease or eliminate reliance on burdensome and potentially biased self-reports to study when, how frequently, and under what circumstances smoking behavior occurs.

Smartphones and, recently, smartwatch technologies have rapidly spread and are widely available[186]. Typical smartwatches house sophisticated sensors that accurately track simple activities, such as step counting. In recent years, methods have been developed that utilize these sensors to detect more complex activities, such as eating and drinking[187, 188]. Previous research[173, 174, 189, 190] has shown the possibility of detecting smoking

using smart devices. However, these studies have employed highly intrusive devices such as respiration bands[174, 191] worn across the chest and two-lead electrocardiographs worn under the clothes to achieve high accuracy in detection. In our previous, laboratory-based work[192, 193] we have shown that smoking can also be detected by leveraging the accelerometer sensor found on a typical smartwatch in conjunction with common machine learning algorithms.

The utilization of smartwatches presents a nonintrusive means of smoking detection that potentially eliminates the need for reliance on self-reporting. The purpose of this study is to extend our previous lab-based work to determine the feasibility and accuracy of our detection method with a population of smokers wearing the device in the natural context of normal, daily activities.

### *5.2.3 Methods*

#### *5.2.3.1 Overview*

Adult smokers were recruited to wear a commonly available smartwatch while recording their daily activities, including smoking and other behaviors that are similar to smoking (i.e., eating, drinking). The data from these recordings were then used in a machine learning exercise to develop an automated gesture detection algorithm. The accuracy of our automated detection was compared against the self-reported information on activities and manual inspection of smoking session data.

#### *5.2.3.2 Recruitment of Participants*

Participants were recruited through flyers, which included study information and a link to an online eligibility survey that was accessible via a clickable URL address and a QR code. The survey asked about participants' smoking behavior, as well as age, gender,

and contact information. Eligibility criteria included: age over 18 years, having smoked at least 100 cigarettes in their life, smoking more than ten cigarettes daily, and preference for smoking with the right hand. The flyers were posted throughout Columbia, South Carolina in areas where smokers were likely to congregate (e.g., coffee shops, bars), as well as online venues such as Craigslist. The incentive for completion of the study was a 100-dollar Visa gift card that was given to each participant after concluding the protocol.

Only participants who met all eligibility requirements were contacted and invited to a study briefing. In the briefing, participants' eligibility was reconfirmed with a smoke CO breathalyzer. A level of 8 ppm was taken as a cut-off, which is slightly higher than cut-off levels of 5-6 ppm suggested for distinguishing smokers from non-smokers in other studies[194-196]. Participants were provided with an Asus Zenwatch and Android smartphone to complete the trial. A 15-minute tutorial was given to each participant on how to use the data collection app and smartwatch and how to fill in the smoking logs, using their smartphone to register the times when they began and finished smoking. Fourteen smokers attended a briefing, two of whom did not meet the criteria of 8 ppm after taking the smoke CO breathalyzer measure and were therefore excluded from the study. Of these twelve participants, data from two were inconsistent and excluded from the analysis because they did not follow the study procedures. In one case, the participant wore the watch on the left hand instead of the right hand and therefore did not collect data from the hand used to smoke. In the second case, large sections of data were missing due to the participant losing Bluetooth connectivity between their watch and their phone by moving more than 30 feet away from the phone. Hence, data from 10 participants were analyzed.

After the study was completed, these ten participants were asked to fill out a brief demographic survey. The survey included basic questions about age, race, ethnicity, gender, and intentions to quit or continue smoking.

### 5.2.3.3 Data Collection and Annotation

The data analyzed in this study consisted of the three-dimensional accelerometer data collected from the Asus Zenwatch (first generation). The accelerometer onboard the Asus Zenwatch is triaxial, and therefore is capable of recording acceleration in three principal axes X, Y, and Z. These three axes are situated on the watch as shown in Figure 5.12, where the Z-axis (in green) is perpendicular to the watch face.



Figure 5.12. An illustration of accelerometer axes on a typical smartwatch is shown.

Although a few apps exist for recording accelerometer data on both Apple and Android platforms, none of them contained the required features, such as recording and transmission of the data to cloud storage or alteration of sampling frequency. Therefore, we developed an app capable of recording, maintaining, and transmitting data to Dropbox as the means of data collection and storage across our cohort of participants. The use of a

customized app allowed for control over the sampling frequency of the data. During this investigation, a fixed sampling frequency of 20 Hz was used.

Each participant was asked to record a total of twelve hours of data over the course of three days. The total of twelve hours was partitioned into seven periods: 4 one-hour periods, 2 two-hour periods and 1 four-hour period. The participants were instructed to schedule these seven periods such that each would contain at least one full smoking session. Due to the large data transfers occurring between the watch and the phone, the battery life of the watch was not able to achieve the full four hours in most cases. In these cases, the participants were asked to record as long as they could until the battery power was nearly depleted.

In addition to the accelerometer data, the participants were instructed to record the beginning and end times of each cigarette in an online logbook using the provided smartphone. A bookmark on the phones linked to a brief Google form that served as their logbook. The protocol involved recording the starting timestamp immediately before beginning a smoking session. In addition, each participant was asked to indicate whether the cigarette was the first from a new pack. After each smoking session, they were asked to report the end of their smoking session as well as the approximate number of puffs during their smoking session.

Smoking sessions were extracted and inspected based on the start and end times recorded in each participant's log entries. The duration of these sessions ranged from 2-20 minutes in length. However, these ranges are, in some ways, misleading. For instance, some of the longer sessions (> 10 minutes) clearly consisted of more than one smoking event. This behavior is typical for chain-smokers but, as per our defined protocol, should

have been recorded as two separate sessions instead of one. Any other gesture that was not within one of the reported sessions was classified as a non-smoking session.

#### 5.2.3.4 A Hierarchical Approach to Detection of the Smoking Gesture

Machine Learning (ML) techniques have been commonly used in the broad field of pattern recognition. Common Machine Learning techniques consist of Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, Artificial Neural Network as well as Rule-based AI, to name a few. In this study, we have integrated Artificial Neural Networks, and Rule-based AI in a hierarchical fashion to improve recognition of smoking activity.

##### 5.2.3.4.1 Artificial Neural Networks (ANN)

In this study two-layer, feed-forward Artificial Neural Networks (ANN)[19] with ten hidden neurons was used as the core engine for detection of smoking gestures. Typically, the creation of an ANN occurs in two main steps: training and validation. Details about the training and validation processes can be found in our previous works[192, 193]. In general, the ANN was trained to produce an output of 1 during the smoking gesture, and a 0 during all other activities. Figure 5.13 provides an illustration of a sample smoking session (with five distinct puffs) and the expected ideal output. In this figure, the patterns illustrated in blue, red and yellow correspond to the X, Y, and Z dimensions of the accelerometer data while the pattern shown in purple denotes the ideal output.



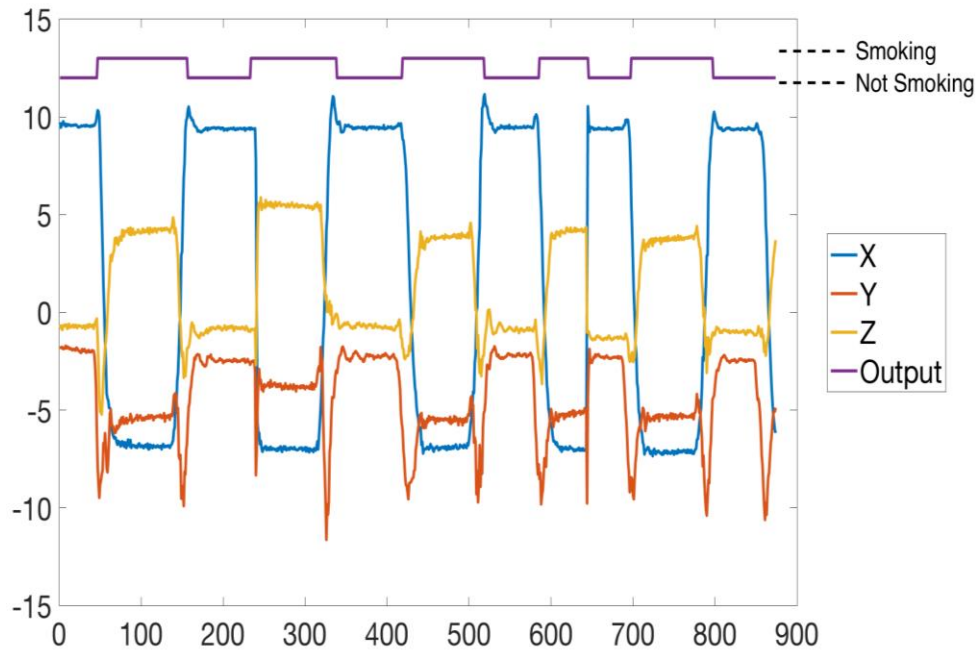


Figure 5.13. An example of a smoking session is shown. Each dimension of the accelerometer data is shown in blue, red and yellow. An ideal output of the ANN is shown in purple where each bump denotes a smoking gesture.

#### 5.2.3.4.2 Rule-Based AI (RB)

Rule-based artificial intelligence constitutes the earliest form of the Machine Learning techniques. RB techniques can be very efficient in circumstances where the actions taken by the AI core can be deduced based on a set of definable rules. The cooperation between the ANN and RB cores can be structured in a variety of ways. In our study, we have chosen a hierarchical model, where ANN operates as the core of the smoking detection, and RB operates in a layer above the ANN. In this arrangement, the RB core is responsible for establishing the beginning and the end of a “puff” gesture, counting the number of puffs, and establishing the beginning and end of a new smoking session. The RB layer also addresses some of the shortcomings of our previous studies[192], where several non-smoking gestures (such as scratching the nose and yawning) caused high numbers of false positives for the ANN. By utilizing the RB layer to establish a minimum

number of puffs within a smoking session, single gestures such as a yawn will be eliminated as a smoking event. The operational directives of the RB core are described later in the paper.

#### 5.2.3.5 Training of the Artificial Neural Network

It is typical to train the ANN on a separate set of data than what is used during the validation step to establish its full functionality (to enforce generalization). This process eliminates the possibility of memorization[197] by the AI. Therefore, the training of the ANN was performed with smoking data collected from 10 volunteers that were not part of the data collection mentioned in the “Recruitment of Participants” section. These volunteers were instructed to use the same smartphone and smartwatch used in the trial to record smoking and non-smoking sessions in a laboratory setting. The training set consisted of 13 smoking sessions that were collected from 7 of the 10 participants and 12 non-smoking sessions from 3 of the remaining subjects. The non-smoking sessions included a variety of activities such as eating (3 sessions), drinking (3 sessions), walking (3 sessions), tying shoes (1 session) and typing on a computer (2 sessions). An example of each gesture is shown in Figure 5.14. Inputs to the network were extracted using a 5-second rolling window, which resulted in a total of 177,450 smoking gestures and 174,080 non-smoking gestures. The smoking gestures were then coded as positive responses and the non-smoking gestures as negative responses. The ANN was trained and validated with this set of data, achieving an accuracy of 95%. Here we define accuracy to be the percentage of correctly predicted smoking and non-smoking gestures.

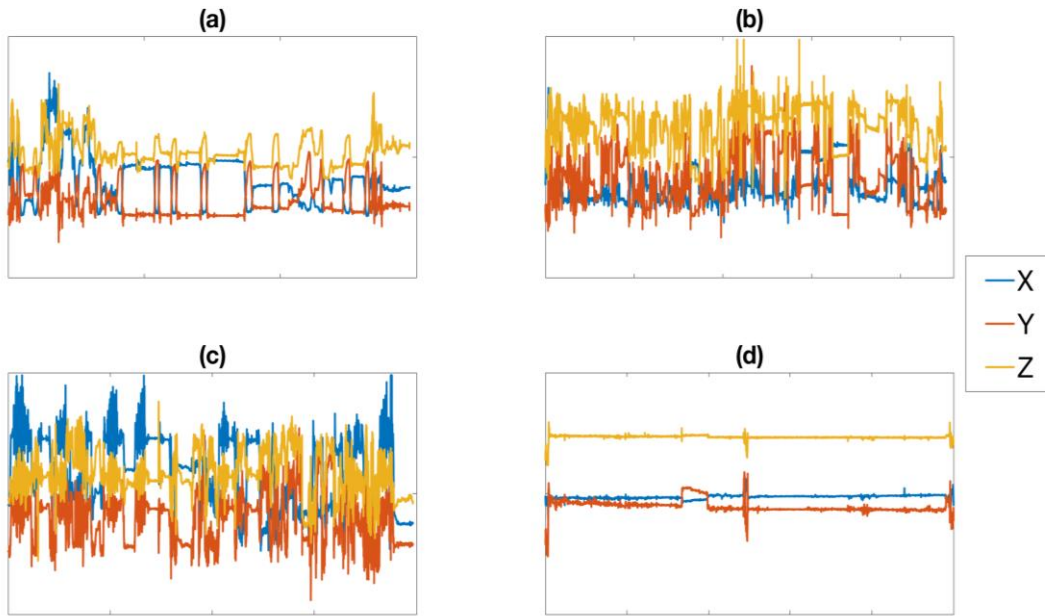


Figure 5.14. Examples of the following non-smoking sessions: (a) drinking, (b) eating, (c) walking, and (d) typing on a computer.

#### 5.2.3.6 Development of the Rule-Based AI from a General Model of Smoking Session

Precise definition of a smoking session is critical for evaluation of a predicted model and development of any rule-based criteria. Development of a template for smoking event is beneficial in a number of ways. First, such a definition can be used to compare the output from our detection mechanism to the actual smoking session recorded by participants. Second, the existence of such a model will help to better define the operating rules of the Rule-Based AI in improving the detection rates.

A smoking session can be defined in terms of its dependent components such as the number of individual gestures and their time dependencies. Figure 5.15 describes the model of smoking that was empirically derived based on our observations of the subjects' data. Based on this model, a smoking session is described by five main parameters: minimum puff duration, minimum and maximum rest time between puffs, maximum

session duration, and the minimum number of puffs per session. A “puff” was defined as the time it takes a person to raise the cigarette to their lips, inhale, and then lower their arm back to the resting position. Therefore, we conservatively define a minimum puff duration consisting of 0.75 seconds (shown in Figure 5.15(a)). Any puff shorter than 0.75 seconds in duration is therefore rejected as a valid puff by the RB AI system.

A minimum of 2.5 seconds and a maximum of 4 minutes were used as the rest time that separates two adjacent puffs (Figure 5.15(b)) belonging to the same smoking session. Two adjacent puffs in violation of the minimum separation criterion were classified by the RB system as the same puff that was incorrectly separated from each other. Correspondingly, two adjacent puffs in violation of the maximum separation criterion are classified to belong to two separate smoking sessions.

Finally, a smoking session was defined to consist of at least 3 puffs that satisfy the above gesture criteria (e.g. puffs must be longer than 0.75 seconds in duration and more than 2.5 seconds and less than 4 minutes from the next puff) and not exceed 8 minutes in duration (Figure 5.15(c-d)). The 8-minute rule was implemented to have a higher precedence over all other rules. A sequence of appropriate puffs that exceed 8 minutes in total length is counted as two separate smoking sessions. This rule has been primarily implemented to address chain-smoking behavior.

In our data, puff- duration never exceeded 5 seconds in length. Therefore, the input to the ANN gesture recognition system consisted of a set of accelerometer data that spanned 5 seconds of observation sampled at 20 Hz (100 points of data). Each set of data included x, y, and z components of the accelerometer, which necessitates an ANN architecture with 300 input points and one output point. The single output of the ANN was

interpreted based on a threshold of 0.85 above which signified a smoking gesture. For more details related to the interpretation of the ANN output please refer to the following work[192, 193].

During the supervised training of the ANN, the onset and offset of the smoking gesture was loosely defined by the supervisor. Loose interpretation of the edge is not consequential since it is a very quick event (in comparison to the gesture itself) and therefore makes very little impact on the duration of a gesture.

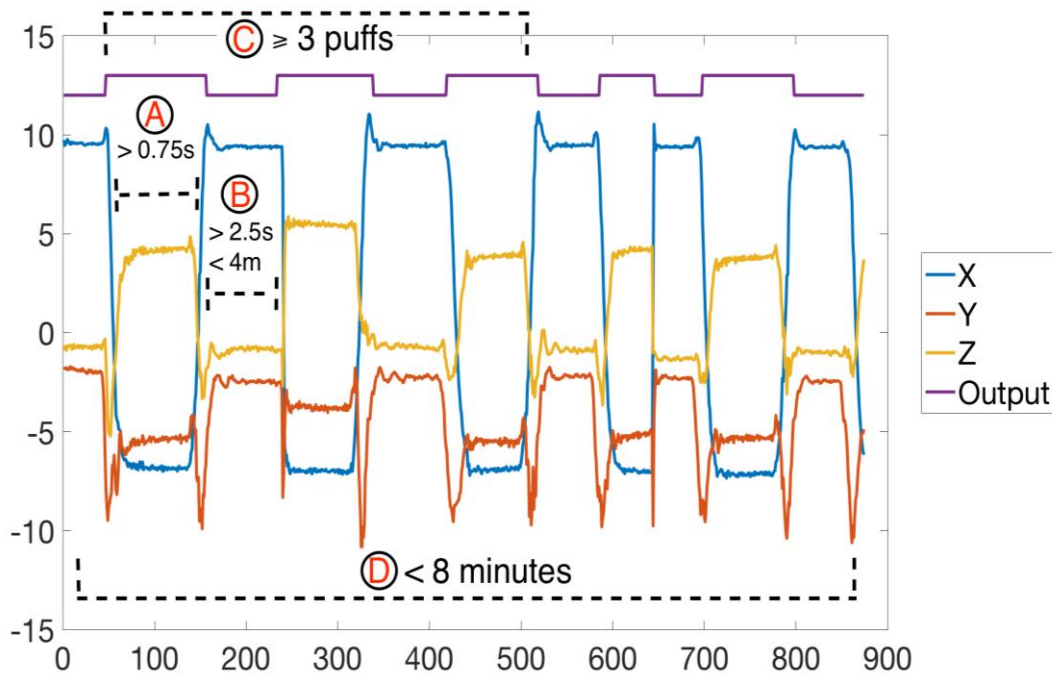


Figure 5.15. Model of a smoking session: a) Puff duration  $> 0.75$  seconds, b) Maximum rest time between puffs  $< 4$  minutes and minimum rest time  $> 2.5$  seconds, c) Minimum number of puffs in a session = 3 puffs, d) Session duration  $< 8$  minutes.

### 5.2.3.7 Evaluation Techniques

Evaluation of automated methods for detection of smoking gestures can be performed at various levels of granularity. At the finest point, every sampled data point (20

points every second) can serve as the subject of evaluation, while at the coarsest point an entire smoking session can be the subject of evaluation. In this work we define our objective as successful detection of each smoking session. The interpretation rules of a smoking session (Figure 5.15) were used to quantify the output of the smoking detection mechanism. The validity of each detected session was established based on comparison to the self-report by the subjects. A detected smoking session was categorized as a True-Positive (TP) if it was corroborated by the timestamps of the self-report and False-Positive (FP) if otherwise. The true positive rate (TPR) was measured using Eq. (5.2.1) where  $S_D$  denotes the number of detected TP sessions and  $S_T$  represents the total number of smoking sessions reported by the participants.

$$TPR = \frac{S_D}{S_T} \quad \text{Eq. (5.2.1)}$$

It is common to provide a measure of False Positive Rate (FPR) to form a more complete evaluation of a predictive system's performance. Calculation of the FPR can be performed based on the Eq. (5.2.2) where  $NS_D$  denotes the total number of non-smoking sessions that were predicted as smoking and  $NS_T$  denotes the total number of non-smoking sessions. However, in this instance proper calculation of the  $NS_T$  term becomes ambiguous. It can be shown that within a 12 hour of recording session a total of 854,400 nonsmoking sessions (of 8 minutes length at 20 Hz of sampling rate) can be extracted via a rolling window. Given that the smoking detection mechanism produced in average two false smoking sessions per participant, the estimated FPR rate would be  $2.34 \times 10^{-6}$ . However, it can be argued that a more meaningful measure of FPR can be achieved based on calculating  $NS_T$  as the total number of contiguous non-smoking sessions (that is the number of 8 minute non-smoking sessions that have no overlap with one another).  $NS_T$  was calculated using

Eq. (5.2.3) where  $H_T$  is the total number of minutes recorded sessions by a given participant,  $WS$  is the window size (in our case 8 minutes) and  $S_T$  is the total number of smoking sessions recorded by the participant. Using this calculation, for a given 12 hour period in which a participant smoked 10 times,  $NS_T$  would be 80.

$$FPR = \frac{NS_D}{NS_T} \quad \text{Eq. (5.2.2)}$$

$$NS_T = \frac{H_T}{WS} - S_T \quad \text{Eq. (5.2.3)}$$

### 5.3.4 Results

#### 5.3.4.1 Summary of Data

##### 5.3.4.1.1 Participant Demographics

Three of the ten participants did not complete the demographic survey. Of the participants who completed the survey, the average age was 32, the minimum age was 27, and the maximum age was 46. There were four females and three males. Six participants were non-Hispanic white, while one was African American. Only one participant indicated that they intended to quit smoking within the next six months.

##### 5.3.4.1.2 Participant Data

In total, 120 hours of data were collected from the ten participants in which 123 smoking sessions were reported. Each data file was first subjected to a low-pass filter to eliminate the high frequency noise caused by movements such as walking or shaking. The effect of the filter can be seen in Figures 5.16 and 5.17. Following the smoothing step, the inputs to ANN were prepared by using a rolling window of 5 seconds.

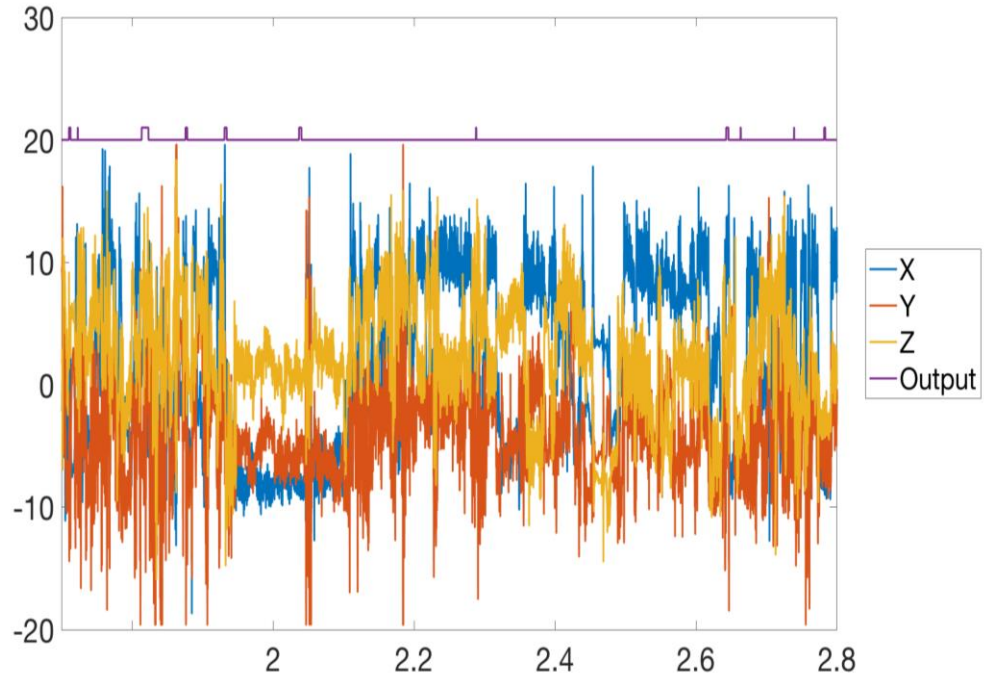


Figure 5.16. A noisy non-smoking session is shown before the smoothing filter with the output of the detection mechanism shown in purple.

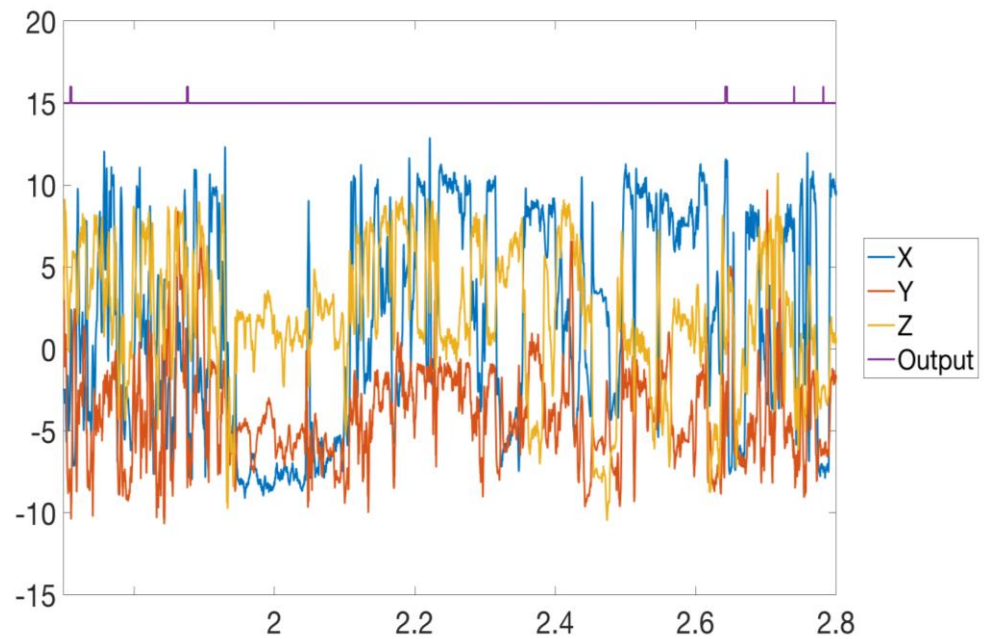


Figure 5.17. A noisy non-smoking session is shown after the smoothing filter with the output of the detection mechanism shown in purple.



Within the 12 hours of recording, participants typically smoked 12 times. On average, the duration of a smoking session was 8 minutes based on the self-report data and 5 minutes based on visual inspection of the recorded sessions. These discrepancies were most likely a consequence of both the additional time required for manual entry in the self-report protocol and human error. Requiring the participants to log their smoking session in an electronic form may have taken some participants a few extra minutes, thus inflating their reported session window.

#### 5.3.4.2 Evaluation Outcomes

##### 5.3.4.2.1 Self-Reporting Accuracy

In total, of the 123 recorded sessions, 27 entries were missing either a start or end time. In these cases, a window of 8 minutes was given preceding an end time with a missing start time or following a start time with a missing end time. Using this metric, the accuracy of self-report (that is the rate of correctly logged smoking entries) was about 78% (96/123). However, it should be noted that we expect the self-report to be lower than the estimated 78%. This expectation is based on close examination of the raw recorded data that would otherwise be impossible to ascertain from self-report data. One such example is shown in Figure 5.18 where the participant did not report a clear smoking session. However, by a close comparison of the recorded session to the model of smoking, one can make a reasonable determination that it is indeed an unreported smoking event. The omission of this session in the log resulted in an increase of the FPR, where it should have contributed to an increase in the TPR. The opposite of this phenomenon has also occurred; that is, a smoking session was reported in a given period yet upon careful inspection no valid smoking event was found in the recorded data (an example is shown in Figure 5.19). If

both of these phenomena were included in the calculation of the self-report accuracy, then it would drop to 71% (88/123 correctly reported sessions).

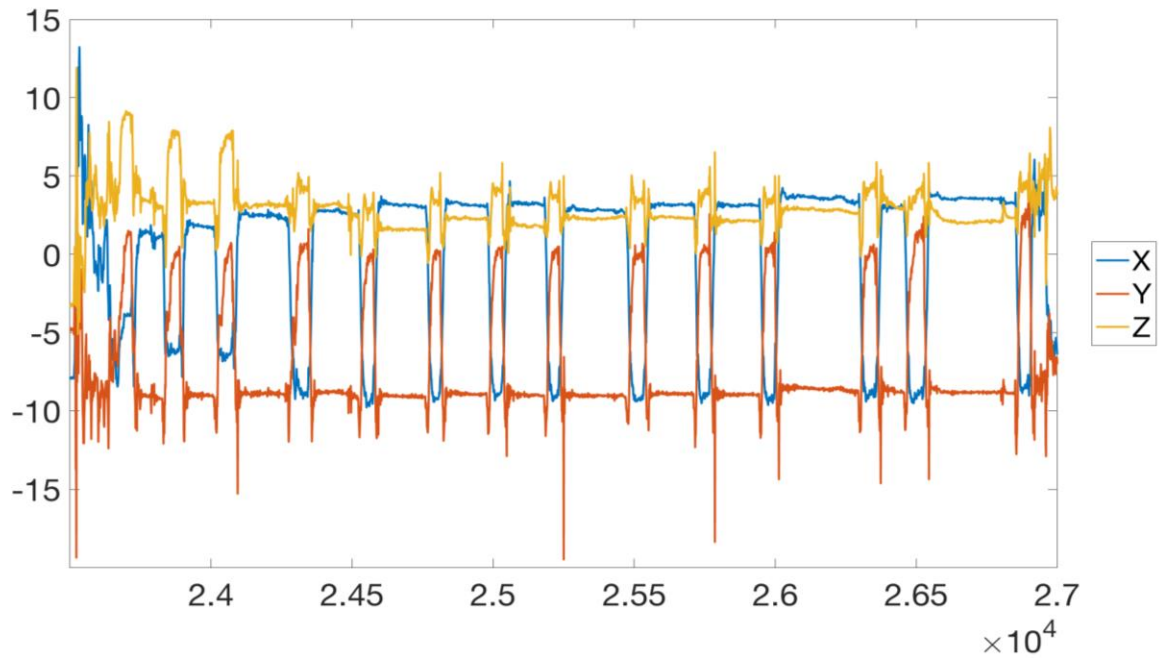


Figure 5.18. This session was not reported by the participant but is an unmistakable smoking session with 13 clear puffs.

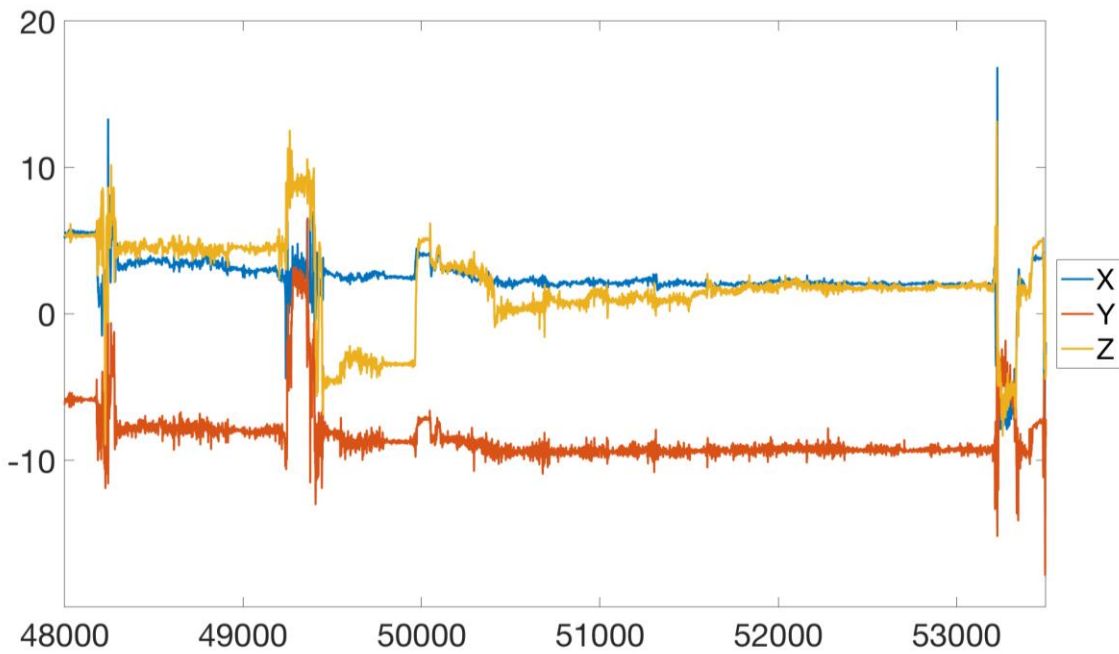


Figure 5.19. This session was reported as a smoking session, but no clear smoking gestures can be identified.

#### 5.2.4.2.2 Detection Accuracy

The results of the automated smoking detection mechanism are shown in Tables 5.2 and 5.3. Surprisingly, the evaluation of the results was not as intuitive as expected. Our initial approach to evaluation (first entry in Table 5.2) was to compare the outcomes of the automated detection mechanism to that of self-reported smoking. However, this approach presumes 100% accuracy of self-report, for which we have cited some contradictory examples (refer to the previous section). If we assume that self-report is 100% accurate, then real errors in self-report (see previous section) lead to underestimating true positive detection rates and overestimating false positive rates. It is therefore paramount to examine and categorize the sources of discrepancy between the two methods. To that end, we define the following categories of discrepancies: “No Smoking,” “Improper Use,” “Abnormal Gesture,” and “True False Negative.” All subsequent investigation of the self-report data was performed by visual inspection of the recorded signals. All detection estimates will be adjusted incrementally as each source of error is eliminated. The modified results are shown in various rows of Table 5.2.

The first category of “No Smoking” denotes no visual presence of a smoking event during the reported smoking period (an example is shown in Figure 5.18). Two such sessions belonged to one participant. These sessions were excluded from the total number of self-reported smoking sessions, which results in a new TPR of 82% (100 out of 121).

The second category, “Improper Use” was one of the biggest contributors in reducing the TPR in this study. “Improper Use” denotes the condition where the participant did not wear the watch as dictated by the protocol of the study (either not on the right wrist or not in the protonated position). This condition can easily be identified and

corrected[193], although the correction mechanism was not implemented and incorporated into this study. A total of 9 sessions were identified via visual inspection to be in violation of proper adherence to the study-protocol and can therefore be excluded from the study. A corrected TPR value of 89% (100 out of 112) can be estimated after elimination of these violations.

The third category, “Abnormal Gesture” denotes the occurrence of smoking gestures that cannot be reproduced in the laboratory settings. These gestures have a clear periodicity that is consistent with smoking behavior but have no other resemblance to our database of smoking gestures. Such conditions may be indicative of smoking in unusual positions such as smoking while lying in the face-down position (possibly from the edge of the bed) or hanging upside down. Various reclined positions, laying down in the face-up position, or lying down on left or right side were investigated without any success in recreating the recorded, anomalous smoking gestures. In future iterations of the detection mechanism utilized in this study, smoking in these positions should be included in our training session of the ANN. However, before retraining the ANN, these curious gestures need to be confirmed as valid smoking sessions and be reproducible in laboratory settings. Depending on whether such gestures can be excluded from this study or not, an upper bound of 99% accuracy can be estimated for the performance of the automated detection mechanism.

The fourth and final category, “True False Negative,” represented the cases where the self-reporting data were correct, and the automated detection mechanism misidentified the sessions. Our thorough investigation identified only one such session. We suspect the abnormally short puffs by this participant as the culprit for this misclassification. The

likelihood of this type of misclassification can be reduced in the future by allowing personalization of the puff duration based on a given person's smoking profile.

Table 5.2. Values for the TPR were calculated by iteratively excluding sessions from the four categories producing false negatives.

Category	Detected Smoking Sessions	Excluded Smoking Sessions	Corrected Smoking Sessions	%TPR
Ground assumption	100	0	123	81%
No Smoking	100	2	121	82%
Improper Use	100	9	112	89%
Abnormal Gesture	100	9-11	101-112	89-99%
True False Negative	100	0	101-112	89-99%

In our evaluation of the FPRs, we faced the same challenges as evaluation of the TPR. A progressive evaluation of the FPR is shown in Table 5.3. Under the simplified conditions (assumption of 100% accuracy in self-reporting), a total of 22 smoking sessions were identified within non-smoking regions of the 120 hours of total recording time. Based on the definition of FPR presented in a previous section, 120 hours of recording time translates into 777 windows of observed non-smoking behavior. Similar to the case of TPR, the following categories of FPR were investigated to understand the nature of the detection mechanism's performance better: "Clearly Smoking" and "True False Positive." The results of classification are summarized in Table 5.3.

Under the conventional technique of assuming 100% confidence in self-reporting data, on average, the detection mechanism achieved an FPR of 2.8% (22 out of 777). However, due to clear presence of errors in self-reports, 2.8% serves as an upper bound estimate of performance, and the actual performance can be expected to be lower than 2.8%.

To obtain a better estimate of the FPR, the first category of “Clearly Smoking” is scrutinized (results are shown in Table 5.3). The “Clearly Smoking” category denotes sessions that at least one smoking event was indisputably present, yet no smoking event was logged during the self-reported period of smoking. An example of the phenomenon is shown in Figure 5.19. A total of 6 such sessions were identified during a careful manual inspection of the recorded data. It is unclear whether such instances should be included in the evaluation of the TPR or FPR. Here we have chosen the latter and have excluded them from the calculation of the FPR. With these excluded the FPR is reduced to 2.1% (16 out of 771).

The second category, “True False Positive”, signifies the cases where smoking detection mechanism performed a true misclassification and, thus, cannot be excluded. A total of 16 such sessions fall into this category. The majority of these sessions contained very jittery and erratic motions, which may be the cause of their misclassification. If so, a more rigorous filtration of high-frequency signals may remove or reduce this category of error in the future iterations of the software.

Table 5.3. Values for the FPR were calculated by iteratively excluding sessions from the two categories producing false positives.

Category	Detected False Smoking Sessions	Excluded Sessions	Corrected False Smoking Sessions	Total Possible Sessions	%FPR
Ground assumption	22	0	22	777	2.8%
Clearly Smoking	22	6	16	771	2.1%
True False Positive	22	0	16	771	2.1%

## 5.2.5 Discussion

### 5.2.5.1 Principal Results

The presented automated smoking detection mechanism demonstrated a conservative TPR of more than 82% for identifying smoking sessions, while achieving a negligible FPR of 2%. Furthermore, the TPR increased to approximately 90% when considering only the smoking sessions when participants adhered to study protocols. Approximately 10 of the smoking sessions were not reproducible in the laboratory session, which will be the subject of future studies to assess how different smoking positions (e.g., while lying down) are accompanied by different gesture patterns or otherwise influence accelerometer readings. Once confirmed as valid smoking sessions, similar gesture patterns can be included in future training sessions of the detection mechanism's underlying ANN. A new TPR can be estimated for the newly trained ANN by assuming 50% successful detection of the anomalous gestures (although, based on the current TPR, 80% is more realistic). A 50% success rate in detecting anomalous gestures will increase the TPR to a 93% accuracy. In contrast, a liberal assessment of the traditional self-report had a maximum accuracy of 71-78%. However, we speculate actual accuracy of self-report may be lower if our analysis of the data from our study is indicative of normal self-report behavior.

### 5.2.5.2 Limitations

There are two primary limitations of the ADSM approach to detection of smoking: technological and methodological. Technological aspects include the battery lifespan, which is of primary interest for applications that require continuous monitoring over waking hours. The wearable device used in our studies (Zen watch) has a limited practical

battery life of nearly 20 hours. However, this battery lifespan may be significantly reduced under high-throughput data exchange conditions, where data are continuously transmitted to another device via a Bluetooth connection. Although a limitation for practical deployment of an automated smoking detection approach, limited battery life can be mitigated in two ways. First, the identification of puffs, smoking gestures, and smoking sessions can be translocated on the watch and therefore eliminate excessive Bluetooth communication. We anticipate a substantial reduction in the power consumption of the smartwatch, returning its lifespan to nearly 10 hours a day. The second mitigation of limited battery life is newly arriving smartwatches with battery lifespan of more than a week. Therefore, the prospect of continuously monitoring smoking behavior for a day or more is highly positive.

A number of methodological issues also limited our current study. The first issue is related to study protocol adherence, which requires participants to wear the smartwatches in a particular fashion (e.g., wearing the watch on the dominant hand). Although these protocols may be acceptable during the early stages of a study, they may be cumbersome during the broader dissemination of this approach. To that end, our existing algorithm should be improved to detect the orientation of the smartwatch (left hand versus right hand, supinated or pronated) either automatically, or during the initial setup stages. Our subsequent work has demonstrated the possibility for automatic correction of the accelerometer data if the watch is incorrectly worn. In addition, study procedures can warn the user if the watch is not correctly worn. The second issue is related to the anomalous and irreproducible smoking gestures we observed. These gestures need to be further



studied, and once confirmed as valid smoking gestures, examples need to be included in future iterations of the smoking detection mechanism.

#### 5.2.5.3 Conclusions

The potential benefits of developing an automated system for detection of human activities are vast. Based on our observations, two distinct conclusions can be stated. First, it is possible to detect of smoking behavior based on tri-axial accelerometer data, and this behavior can be distinguished from other similar gestures. Second, an automated smoking detection approach to study of smoking behavior may be substantially more reliable than approaches that rely on tradition self-report. Third, with an accurate, automated system in place, reliance on self-reporting could be eliminated, thus decreasing the burden on a participant without losing any benefits. The resulting data collection system could allow for a range of unobtrusive studies of how context, including that which can be captured by GPS systems, influences smoking behavior, targeted surveys around smoking events, and targeted communications for those who are trying to quit. Furthermore, this automated system may easily be expanded to detect increasingly popular electronic cigarette smoking, for which behavioral gestures accompanying consumption are very similar to cigarette smoking but for which the patterns of behavior and their context are much less well understood.

### 5.3 Clinical Quantification of Smoking Topography Using Smartwatch

#### Technology

##### 5.3.1 Abstract

While there have been many technological advances in studying the neurobiological and clinical basis of tobacco use disorder and nicotine addiction, there have

been relatively minor advances in technologies for monitoring, characterizing and intervening on smoking behavior. The use of emerging smartwatches can help to better study temporal patterns and topographies of smoking in natural settings. In this study we compared the results of characterizing smoking topography using a novel mobile software with pocket CReSS and video recording. Adult smokers (N=27) engaged in a video-recorded laboratory smoking task. Participants smoked a cigarette using pocket CReSS to assess smoking topography while also wearing a Polar M600 smartwatch. An in-house software, ASPIRE, was used to record accelerometer data to identify the duration of puffs and inter-puff intervals (IPI). Agreement between staff-observed, CReSS-recorded, and ASPIRE-recorded smoking behavior was examined. ASPIRE produced more consistent number of puffs and IPI durations relative to CReSS, when comparing both methods to visual puff count. After filtering implausible data recorded from CReSS, ASPIRE and CReSS produced consistent results for puff duration ( $R^2 = .79$ ) and IPIs ( $R^2 = .73$ ). Agreement between ASPIRE and other indicators of smoking topography was high, suggesting that the use of ASPIRE is a viable method of passively characterizing smoking behavior. Moreover, ASPIRE was more accurate than CReSS for measuring puffs and IPIs.

### *5.3.2 Implications*

Results from this study provide the foundation and support for utilizing ASPIRE to passively and accurately monitor and quantify smoking behavior in situ. Better understanding of real time smoking behavior can be helpful in numerous applications including initiation of just-in-time intervention mechanisms. Furthermore, automated recording of smoking, or other related human activities, removes the burden and recall biases of self-reporting.

### 5.3.3 Introduction

Tobacco use disorder (TUD) is the leading preventable cause of death worldwide (including the US) [164] and the costs associated with its treatment and prevention remain a major economic burden on society [198]. Therefore, a better understanding of the behavioral elements and mechanisms that maintain smoking behavior is critically important for preventing future smoking-related illnesses. While there have been substantial technological advances in studying the neurobiological and clinical basis of TUD and nicotine addiction, there have been relatively minor advances in technologies for monitoring, characterizing and intervening on smoking behavior. Therefore, there is a critical need for leveraging emerging technologies that may help provide personalized strategies for smoking cessation.

Traditional approaches to the study of human behavior primarily rely upon self-reporting or laboratory observations. Despite strengths found in self-report and laboratory-based research, those techniques, by design, are prone to having limitations in external validity. To more fully characterize and understand factors influencing people's behavior, enabling technologies must be developed to allow non-intrusive and longitudinal observation of human behavior in natural settings. Mobile devices are well-positioned to passively assess a person's behavior in the aforementioned context.

Mobile devices contain a rich array of sensors (accelerometer, gyroscope, magnetometer, barometer, GPS, heart rate, ECG, oximeter) and may serve as a powerful platform for capturing and studying human behavior. In addition, the availability of mobile devices is an international phenomenon and their use is not confined to any particular socioeconomic class. Therefore, the use of smart and wearable devices for delivery of

sensing behavior has the potential to passively observe health behaviors and be deployed internationally without confinement to any socioeconomic, political, or geographical barriers.

In recent years, there have been several reports of utilizing commercially available smartwatches in studying human activities. These include generic activities such as step counter, sleep detection, and rest periods, while others include more specific activities such as eating[199], drinking[200], or smoking.[175] Previous work has established the use of wrist-worn devices in observing and interpreting smoking behavior in laboratory settings[192, 193] and in situ[175, 201]. Some of these devices use proprietary sensors[173, 191, 200, 202], while others use off-the-shelf devices such as smartwatches[175, 176, 192, 201, 203]. The use of smartwatches in continuous monitoring of human activities and behavior is compelling for several reasons including, their availability, cost, popularity; and the convenience and completeness of data collection. The data connectivity that is afforded by smartwatches adds a critical component to their appeal, allowing for real-time observation and interpretation of activities that can lead to an immediate deployment of the appropriate intervention. AI-assisted detection of smoking with smartwatch technology[175, 176, 189, 192, 193, 201, 203] can reduce the burden of its operation by the user, team of research scientists, and the caregivers. Despite their potential, the accuracy and resolution of data collected by smartwatches has not been systematically explored in comparison to traditional approaches to the study of smoking behavior. More specifically, while it has been shown that detection of smoking is possible with a smartwatch[175, 192, 201, 203-205], the use of smartwatches in better exploring smoking topography of human subjects remains unanswered.

Observation and automated detection of smoking using smartwatches also has several compelling aspects. First, smartwatches consist of sufficient storage capacity to record and store sensor data for a duration longer than 24 hours. The recording duration can be extended into months by the addition of micro SD storage media. Second, the collected data will require no additional action by the user or the participant of a study, qualifying this method of data acquisition to be highly unobtrusive. Third, unlike self-reporting approaches, continuous recording of sensor data can provide a comprehensive report of a person's activities in natural settings. For instance, proper interpretation of the sensor data can provide a detailed view of related human activities such as drinking, eating, sleeping, exercising and smoking all in one experiment. The collection of such detailed ensemble of activities will be nearly infeasible through the use of self-reporting when observed in situ. Fourth, the real-time connectivity of smartwatches allows for real-time observation of human behavior, which can be used in numerous ways to study or augment human behavior. For example, the adherence of a subject to study protocols can be viewed and confirmed, and if necessary, notifications and reminders can be sent to the participants. Real-time and continuous connectivity with participants allows for the initiation of the appropriate actions, paving the way for personalized intervention or cessation approaches. Finally, the automated detection of activities (such as smoking) can initiate timely questionnaires, actions, or notify the appropriate members of the participant's social network.

In this report, we present an evaluation and comparison of the quality of smoking data collected by smartwatches, CReSS, and video recordings of human subjects in a laboratory setting. In particular, we compare and contrast the accuracy of observing smoking

topography using the Automated Smoking Perception and REcording (ASPIRE) smartwatch application and the Clinical Research Support System (CReSS) device. We resort to human annotation of visual recordings of smoking sessions when possible to resolve substantial disagreements between the ASPIRE and CReSS approaches. We also explore the additional capabilities of the ASPIRE-based method and comment on the significant advantages that it affords and its novel future utilities.

#### 5.3.4 Methods

##### 5.3.4.1 CReSS Device

Clinical Research Support System for Laboratories (CReSS Pocket; Borgwaldt KC, Inc, Richmond, VA; <https://www.borgwaldt.com/en/products/smoking-vaping-machines/smoking-topography-devices.html>) is widely used for studying smoking in a laboratory setting [206-209]. Though the CReSS devices provide objective measures of smoking topography and is amenable to use a laboratory setting, it is relatively expensive (~\$5,500), and interferes with the natural smoking experience. Those issues limit its utility for characterizing ad lib smoking in a smoker's natural environment.

##### 5.3.4.2 Characterization of Smoking Topography using Smartwatch

A smartwatch-based method allows participants to smoke freely in their natural settings without the need to use an intermediary device. Our previous work reported the development of an Android Wear OS-based software (ASPIRE) package that is capable of recording[192], and automated detection of smoking gestures (puff)[201]. ASPIRE incorporates a hierarchy of AI techniques in order to achieve automated detection of smoking sessions with as high as 97% success in laboratory settings[192, 203], and 90% success in natural settings[201].

In the current study, participants were provided a smartwatch (Polar M600) and Android smartphone for the duration of data collection. The ASPIRE app was downloaded to both the watch and the phone. The smartwatch app listens and collects the data from the participant and then sends it via Bluetooth connection to the smartphone app. The smartphone then uploads the data to a secure server for storage. The current version of ASPIRE can be obtained from the corresponding author and installed in either a phone/watch pair or a stand-alone app on the watch.

#### 5.3.4.3 Data Collection Protocol

Participants were first outfitted with a Polar M600 to wear on their left hand. Participants were asked to follow a prompt screen on a computer in the laboratory that gave them precise instructions on what behaviors to perform as well as how long to execute each behavior. An overall view of the experimental paradigm with associated durations is shown in the supplementary section, Figure 5.20. For this study, participants were asked to smoke a total of six minutes, in which they were asked to split evenly between their left and right hands. In addition to smoking, they were also asked to record over seven minutes of other movements including 52 seconds of “packing” their cigarette package. These additional gestures will be excluded for the purposes of this study but will be critical in future work to further validate the specificity of ASPIRE.

Condition	Baseline	Natural Movements	Use Phone	Drink Bev.	Pack Cigarettes	ad lib Smoking (CReSS)	Drink Bev.	Use Phone	Natural Movements	End
Duration	16 sec	220sec	62 sec	52 sec	52 sec	397 sec	52 sec	58 sec	220sec	5sec
Timeline (sec)	0-16	16-236	236-298	298 - 350	350 - 402	402 - 799 (6.00 min of actual smoking)	799 - 851	851 - 909	909 - 1129	1129 - 1134
						1 <sup>st</sup> 3 min. = L. Hand	2 <sup>nd</sup> 3 min. = R. Hand			

TOTAL TIME: 1134sec (18:54)

Figure 5.20. Outline of the protocol used for collection of in laboratory data.

The CReSS device was used to record the following measures: puff volume, average flow, peak flow, time of peak flow, duration, and inter-puff interval (IPI). The two measures of interest for this study are puff duration and IPI. The puff duration is the length of time in milliseconds that a person inhales for a given intake. The IPI is the number of milliseconds between the end of one puff and the beginning of the next. CReSS records both a high-level and a detailed view of these measures. A median measure is used for the high-level view due to the small sample size and existence of outliers that are inherent to the device. The detailed view contains information about each one of the measures per puff. In addition to the data collected by the CReSS and smartwatch devices, videos of each session were also recorded, coded by two independent raters (inter-rater reliability =1.0).

#### 5.3.4.4 Participants

Participants were recruited via community advertisements, attended an in-person screening visit to determine eligibility and then an experimental smoking visit. Participants gave written informed consent approved the Medical University of South Carolina IRB, and received financial compensation for study participation. Inclusion criteria were: being age 18 years or higher, having an expired carbon monoxide (CO) concentration of  $\geq 6$  ppm



(to confirm smoking status) and being willing and able to comply with protocol requirements. The participants (N = 35; female = 13) were on average 43.91 (+/- 12.76) years old with a CO level of 26.57 (+/- 12.32) ppm. Due to technical or recording errors, 8 participants had incomplete CReSS (n=5) or ASPIRE (n=3) data, resulting in a final analytic sample of N=27.

#### 5.3.4.5 Data annotation procedure

The first step in the evaluation procedure was to annotate the data collected from smartwatch, CReSS, and video recordings. Due to the time and effort required for the video recordings, only the puff count from each hand was visually annotated. The annotation for the smartwatch and CReSS device consisted of marking the timestamp associated with the beginning and end of each puff. Using this information, puff duration and Inter Puff Interval (IPI) can be measured. In order to isolate the portion of the data that contained smoking, the timestamp of each data point was referenced to the design protocol. Then a well-trained researcher recorded the total number of puffs in the region as well as the start and end of each puff. Each puff is easily identifiable as first starting with a slight or negligible change in the x dimension ( $\pm 1$ ), a moderate decrease in the y dimension (-4) and a sharp decrease in the z dimension (-8) from a resting position. This is then followed by a period of uninterrupted and equilibrated values of x, y, and z at around  $9 \text{ m/s}^2$ ,  $-5 \text{ m/s}^2$  and  $-3 \text{ m/s}^2$  respectively. The end of a smoking gesture was then marked as the return of the x, y and z values to a “resting” state. These numbers vary per participant but will follow the same general shape.

#### 5.3.4.6 Evaluation and Exclusion of Data

The first step in the evaluation of this work was to annotate the data collected from the participants' puffs (protocol described in section E). The puff duration and IPIs were calculated using these annotations. In the case of ASPIRE, the known sampling rate of 30Hz was used to convert the timestep units into millisecond units. Figure 5.21 shows an example of a full set of data (shown in upper right box) recorded by ASPIRE, and a subsection of that data that contained smoking (annotated puffs are denoted with asterisks). The period of smoking shown in Figure 5.21 accounts for approximately three minutes of data.

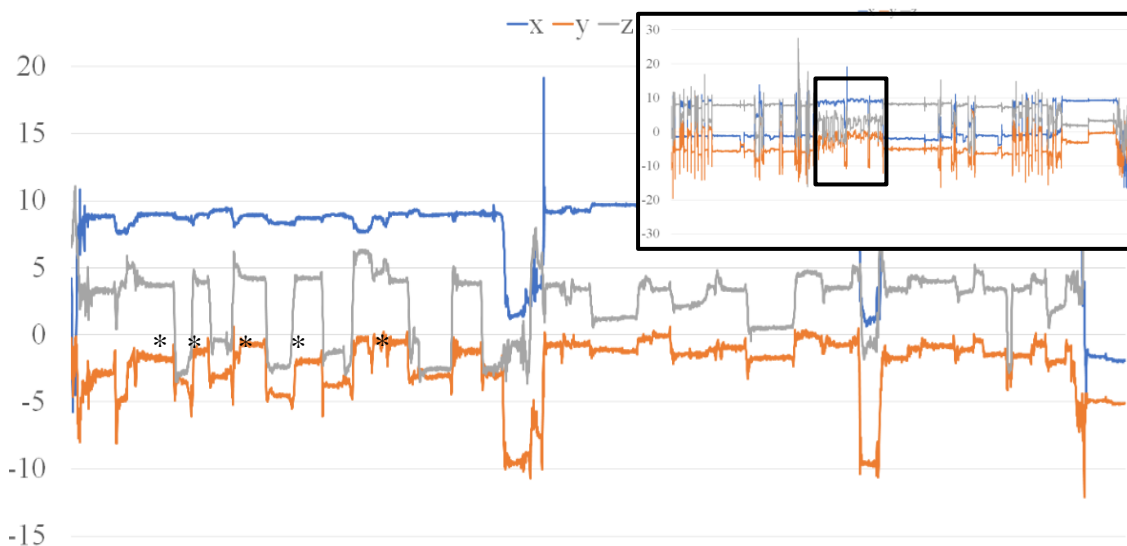


Figure 5.21. A sample of ASPIRE's recording session illustrated in the upper right corner. The main and larger figure depicts the portion of the image that corresponds to a smoking session (asterisk indicate the start of a puff).

The second phase of the evaluation consisted of calculating the Pearson correlation coefficients between the extracted puff durations and IPIs reported by CReSS and ASPIRE. Correlations were first examined using the reported median data for all subjects and then followed by analysis of the data for each individual subject.

Due to the switching of the smoking hand without the transfer of the smartwatch, half of the smoking session was not recorded by the smartwatch. Therefore, using the video recordings of the smoking sessions, we limited our comparison exercise to the portion of the smoking sessions that were recorded by both CReSS and ASPIRE. In addition, in some instances CReSS reported implausible measures, inclusion of which would provide an inaccurate comparison of the three methods. For instance, CReSS reported puffs with a duration of 5ms or IPIs of over one minute at the beginning of each smoking session. The long IPI at the beginning of the smoking session is the time the device was turned on to the time of the first puff and were therefore removed from our analysis. The implausibly short puffs reported by CReSS can be explained by a participant performing rapid and multiple puffs such that neither ASPIRE nor the video recordings could identify. In such instances we report results with and without the included implausible data since they serve as clear demonstration of some nuances of the CReSS device.

### *5.3.5 Results*

#### *5.3.5.1 Overview of the Collected Data*

Each accelerometer data file collected by ASPIRE contained 20 minutes of data, which is consistent with the experimental protocol. As a first step in our comparison of the two methods, histograms were created for individual puff durations and IPIs combined across all subjects (shown in Figure 5.22 and Figure 5.23). The blue bars in these figures correspond to the values produced by the CReSS device and orange bars correspond to the values produced by ASPIRE. Although the distributions of the puff durations were very similar between the two methods, the distributions of IPI values were different (shown in Figure 5.23). While the two histograms demonstrate a general agreement, they differ

notably in reporting the number of small IPIs. For the CReSS data, there is a spike in very low values corresponding to the IPI value of 0-1 second. The median of the IPIs reported by CReSS in this range was 0.33s.

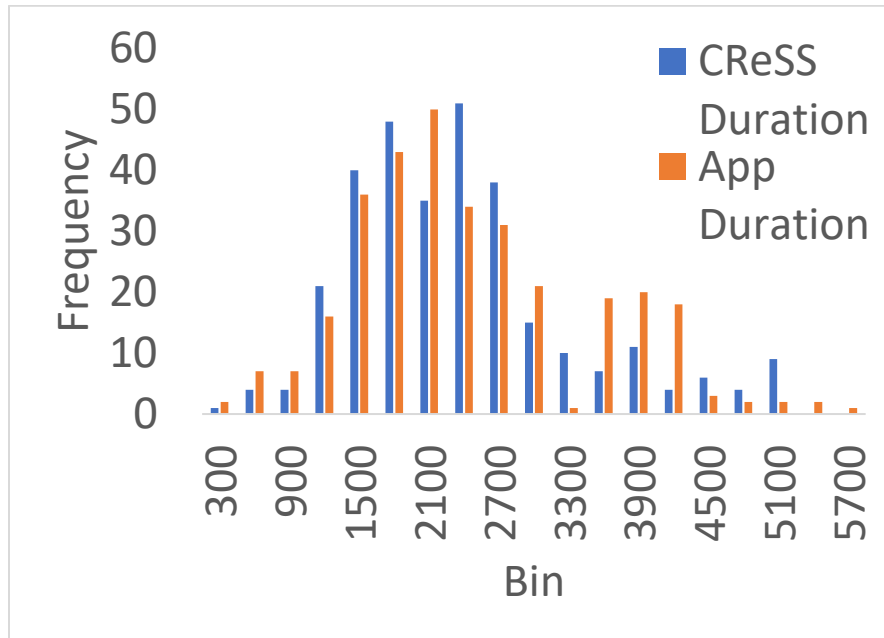


Figure 5.22. Comparison of individual puff durations collected via CReSS (blue) and the App (orange).

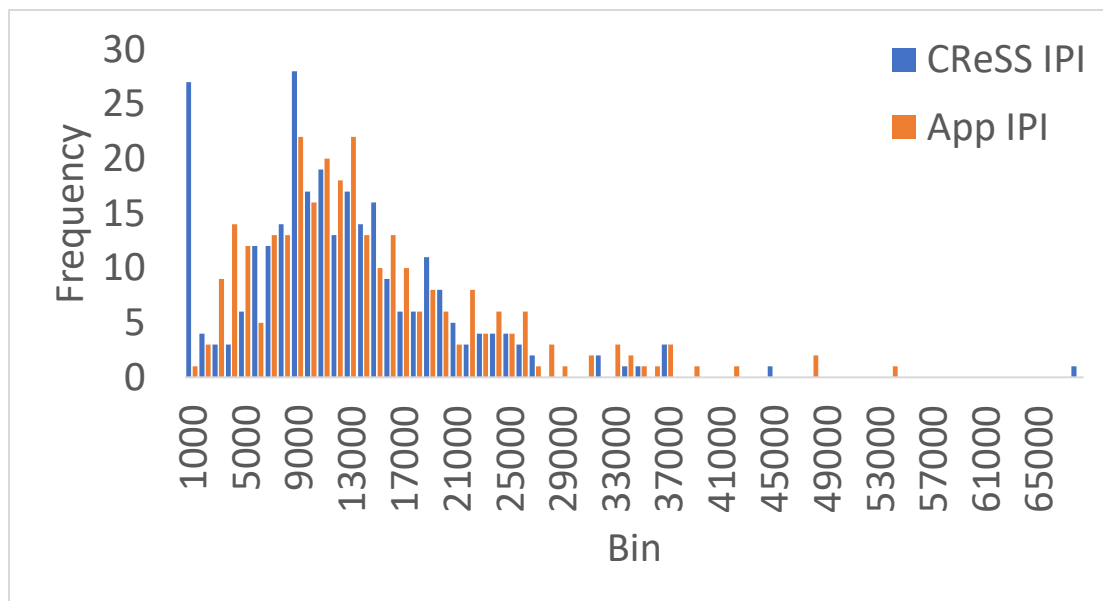


Figure 5.23. Comparison of individual IPIs collected via CReSS (blue) and the App (orange).

### 5.3.5.2 Comparison of Overall Statistics

The number of puffs that the CReSS device recorded was compared to a visual inspection of each participant's respective video recordings. In 80% of cases, the visual puff count and the count reported by CReSS were within  $\pm 2$ . However, there were four cases where the puff counts differed by as much as  $\pm 6$ . Figures 5.24a and 5.24b show the correlations between the overall visual puff counts for the left hand of each participant compared to the puff counts reported by ASPIRE and CReSS (respectively). The  $R^2$  value for the visual puff count and the counts reported by ASPIRE was 0.79 whereas the  $R^2$  between the visual puff count and the CReSS reported count was 0.52. The participant that caused the most deviation in both comparisons was P14 (in red in Figures 5.24a and 5.24b) with visual, ASPIRE and CReSS reported counts of 14, 8, and 38 respectively.

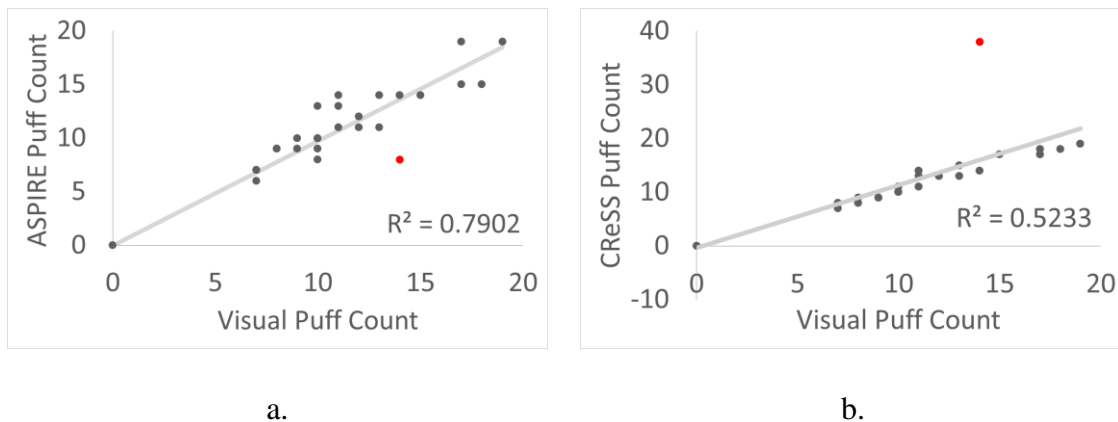


Figure 5.24. Comparison of the visual puff count versus the a.) ASPIRE puff count and b.) CReSS puff count. In both figures' participant P14 was an outlier and is colored red.

The  $R^2$  between the median puff duration, and median IPI across all patients is illustrated in Figure 5.25a and Figure 5.25b.  $R^2$  values of 0.7926 and 0.7309 ( $p < 0.001$ ) were calculated for the median puff durations and median IPIs respectively, indicating a high level of correlation between the data reported by the two methods.

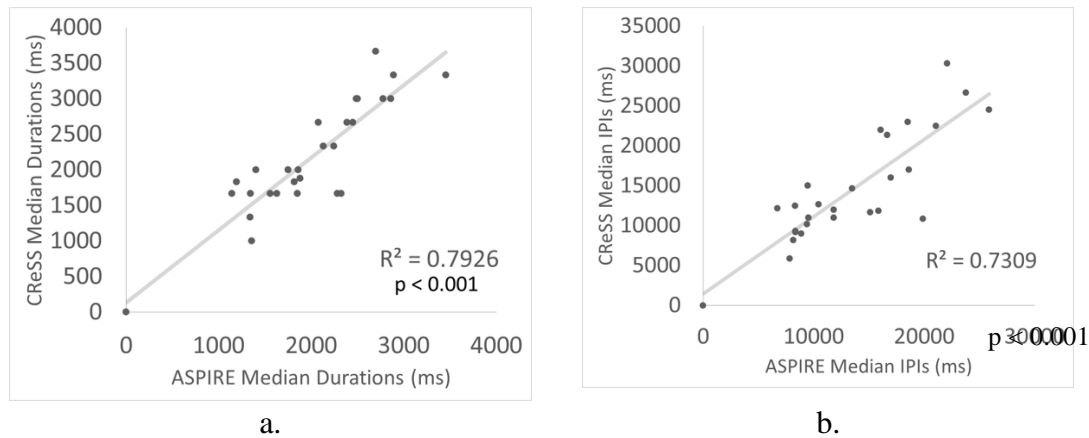


Figure 5.25. Comparison of the overall a.) CReSS reported median puff duration v. ASPIRE reported and b.) CReSS reported median IPI and ASPIRE reported.

### 5.3.5.3 Comparison of Individual Puffs

Due to the enigmatic nature of smoking topography data acquired by CReSS, the data is not usually used to perform detailed statistical analyses or inferences and instead the median values for puff duration and IPI are used. However, availability of the more reliable data from ASPIRE allows for the meaningful study of mean, standard deviation and other statistical moments of puff duration and IPI for each participant. Figure 5.26 illustrates the smoking topography for a representative participant (P2) reported by CReSS and ASPIRE. Both methods reported a total of ten recorded puffs separated by nine IPI intervals. Also, both methods report similar values for the median puff duration and IPI. However, it is clear that the duration of the entire event is highly discrepant across the two methods. Furthermore, visual inspection of the smoking session reported by CReSS consists of only eight decipherable puffs (green regions). This is due to very short IPIs of 5 milliseconds that render two puffs invisible in the figure. On the other hand, the same smoking session reported by ASPIRE is well organized into the expected shorter puffs that are separated by longer IPIs. The trend marked as cCReSS in this figure, corresponds to

the corrected CReSS data by only correcting five elements (puffs or IPIs out of 20) of the smoking session. The correlation between the CReSS and ASPIRE reported data improved from 0.05 to 0.8 after correcting for the discrepant data. Figures 5.27 and 5.28 demonstrate other examples of similarity between the CReSS and ASPIRE data after correcting for outliers. These examples have  $R^2$  values in the ranges of 0.77-0.83 for puff duration and 0.97-0.99 for IPI. These correlation values indicate a similarity reported by the two different methods with statistical significance of  $p < 0.005$ .

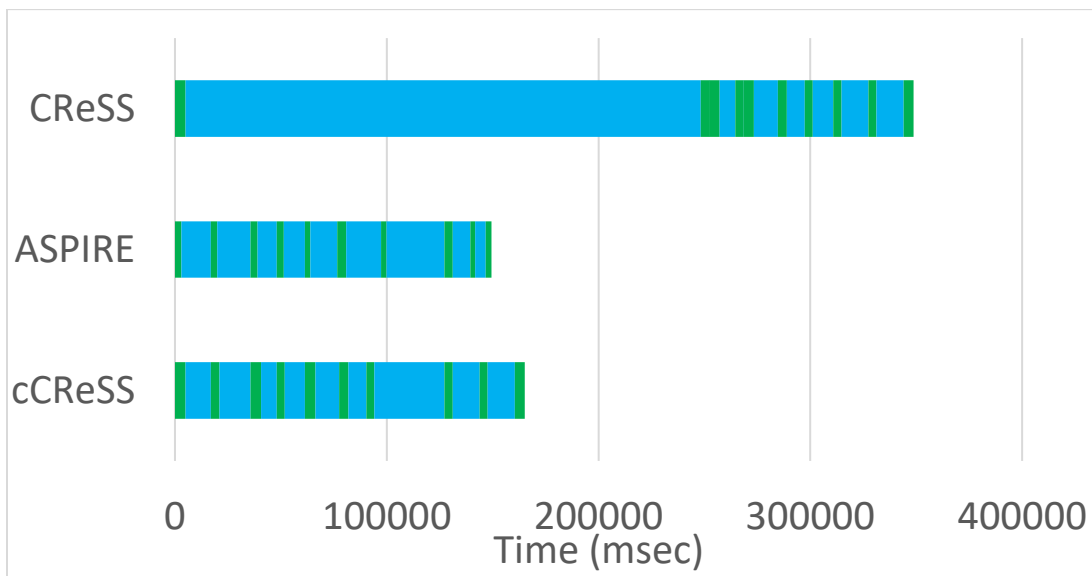
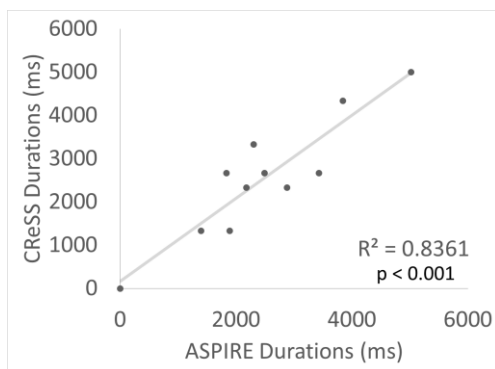
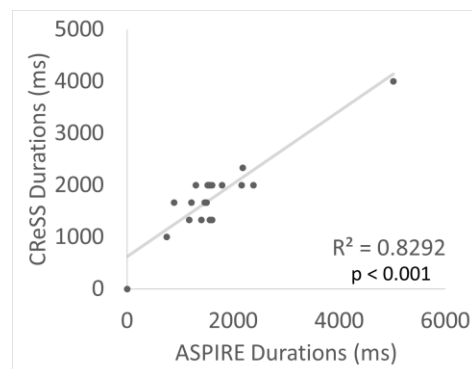


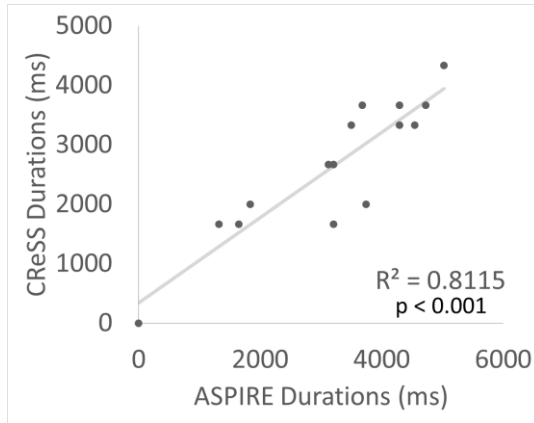
Figure 5.26. An illustration of smoking topography reported by CReSS, ASPIRE, and corrected CReSS. The puff durations and IPIs are illustrated in green and blue respectively.



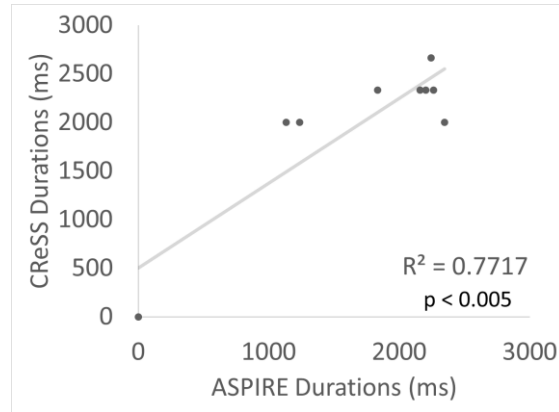
a.



b.

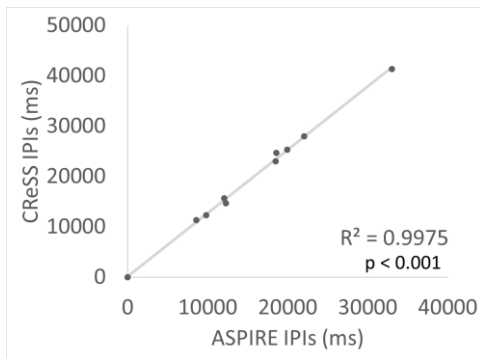


c.

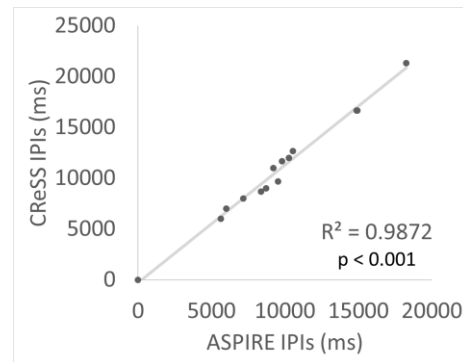


d.

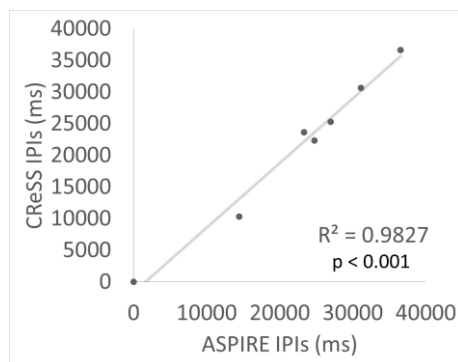
Figure 5.27. Correlations of individual puff durations collected via the CReSS device and ASPIRE for participants a.) P15, b.) P17, c.) P19 and d.) P8.



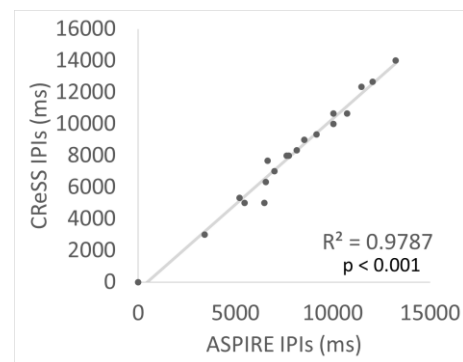
a.



b.



c.



d.

Figure 5.28. Correlations of individual IPIs collected via the CReSS device and ASPIRE for participant a.) P15, b.) P19, c.) P7 and d.) P17.



#### 5.3.5.4 Preliminary Results for Puff Volume

One critical advantage of the CReSS device is its capability in collecting volumetric data related to smoking by measuring the volume of air that is inhaled with each puff of a cigarette. In this exercise we explored the possibility of inferring the volumetric information from puff durations. Figure 5.29 shows preliminary comparison of the median puff durations recorded by ASPIRE versus the median puff volumes reported by CReSS. These two measurements exhibit a  $R^2$  value of 0.5216, which indicates a clear evidence of a relationship between these values with a statistical significance of  $p < 0.001$ . The correlation between puff duration and volume may be further improved by considering other factors, such as the participant's height and weight.

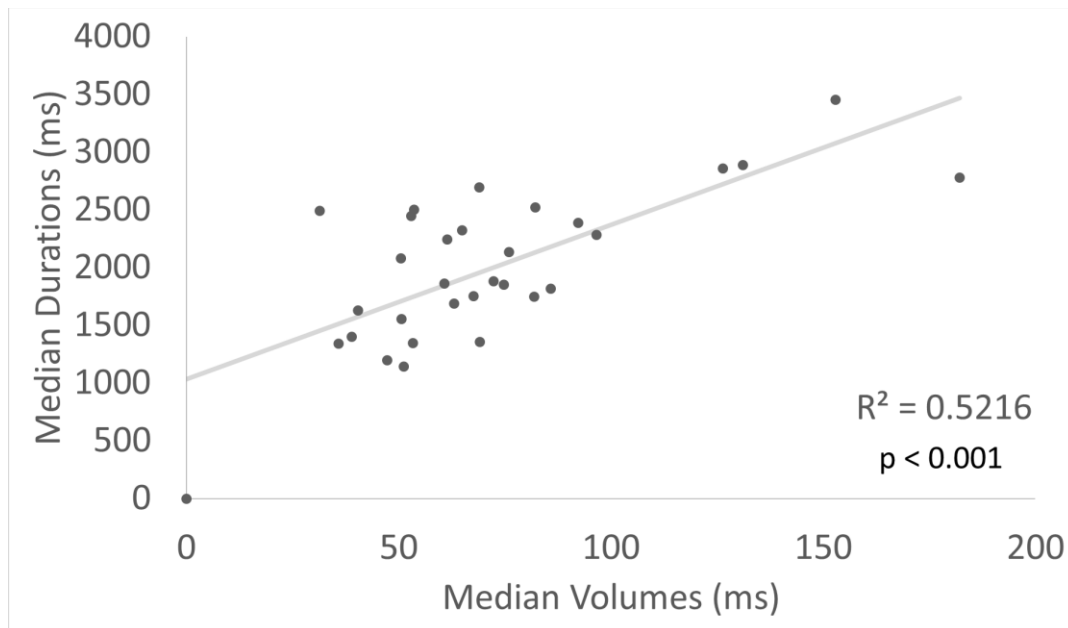


Figure 5.29. Correlation of median puff volume and median puff duration.

#### 5.4.6 Discussion

Findings from this proof of concept study provide direct evidence to support using ASPIRE to passively collect multiple components of smoking behavior, findings which

pave the foundation for basic smoking research and translation for clinical interventions to naturalistic settings. The study revealed three important findings. First, ASPIRE is highly effective in accurately detecting when a person initiates smoking and how many puffs of a cigarette is inhaled. Secondly, ASPIRE may be used to accurately characterize the duration of each puff—which may provide a dose indicator when used in conjunction with puff count. Finally, ASPIRE can detect time between each puff (inter-puff interval), which may provide a meaningful metric of episodic smoking compulsivity.

Smartwatches have the potential to significantly advance the study of human behavior in situ, however the reliability of the information reported by wearable devices has been questioned. In this study we have investigated and demonstrated the reliability of the data reported by ASPIRE in the laboratory settings compared to the CReSS and visual recording of the smoking sessions. Moreover, though other technologies have been recently developed to detect smoking behavior at the puff or session level[175, 176, 201], this is the first study to demonstrate characterization of smoking topography (namely duration and IPI) as performed by ASPIRE.

The ability to passively collect and accurately characterize multiple components of smoking behavior in the natural environment is a critical step in monitoring smoking, characterizing smoking outcomes in outpatient clinical trials, and developing real-time adaptive interventions/personalizing smoking cessation interventions. Much of our knowledge about the mechanisms that elicit smoking behavior is obtained from observation of behavior in laboratory settings. For example, research has examined how exposure to smoking stimuli (for example, image of cigarette lighter) [210], acute stressors or mood inductions [e.g., 211, 212], fasting [213], and interventions [214, 215] affect smoking

behavior under controlled laboratory conditions. ASPIRE can be used to examine whether laboratory-based findings generalize to real-world settings. In future studies, ASPIRE-detected smoking can be compared to randomly prompted app, text, or smartwatch surveys asking about stress. Alternatively, several passive technologies can be combined. For example, measures of electrodermal activity or heart rate (i.e., physiological arousal) recorded via mobile devices can be collected in addition to assessing the timing, count, and characteristics of cigarettes smoked in real-world settings.

In addition, real-time technology has the potential to greatly improve assessment of smoking outcomes in smoking cessation clinical trials [216]. The traditional outcome measure in smoking cessation clinical trials is biomarker-confirmed abstinence [217, 218]. Typically, this is assessed via participant report of abstinence for a certain number of days (i.e., 7-day abstinence) and confirmed in laboratory via carbon monoxide or cotinine. These outcomes are subject to errors in retrospective recall and intentional misreporting. ASPIRE can provide objective evidence of smoking while also indicating when the smartwatch is removed for the purposes of determining adherence with wearing the smartwatch. Remote technologies also help to extend the reach of clinical trials. Individuals with the necessary technology can participate in a smoking cessation trial remotely (i.e., at a different location than the research team) while still providing rigorous evidence of smoking and/or abstinence.

Finally, the incorporation of AI and Machine Learning techniques to automatically detect and report smoking behavior can assist in the delivery of personalized and Just-in-Time interventions. AI algorithms can incorporate the precise information regarding the context and timing of cigarettes smoked gained from ASPIRE to determine when an

individual is most likely to smoke. Interventions can be delivered pre-emptively to prevent smoking or relapse [e.g., 219].

A number of limitations must be considered when interpreting the current findings. First, validation of ASPIRE requires comparison to a gold standard measure. However, CReSS, as the gold standard comparator used in this study, suffers its own limitations. The CReSS device identifies smoke topography only based on the inhalation patterns and does not incorporate any information regarding the exhalation activity. Therefore, the CReSS device may identify a single puff that is composed of numerous discontinuous puffs as multiple short puffs separated by short IPIs. For instance, in Figure 5.23 it was shown that CReSS recorded a significant number of IPIs of 1 second or less. While IPIs of this length are theoretically possible, their appearance in such an abundance reported by CReSS is highly suspect. The short IPIs reported by CReSS contribute to skewing the overall statistics presented which lowered concordance between the CReSS puffs and the ASPIRE collected puffs. Based on the review of the visual recordings in this study, we have confirmed that the ASPIRE approach provides a more consistent and reproducible report of the smoking topography than the CReSS device. Furthermore, we have also demonstrated the consistency of smoking topography reported by smartwatches and the CReSS device in laboratory settings. Our results conclude the puff duration and IPI reported by both devices exhibit a substantial degree of correlation after the exclusion of the outliers reported by the CReSS device.

A second limitation of this study is that the laboratory is an unnatural smoking environment that may elicit unnatural smoking behavior from the participants. ASPIRE can record and report smoking behavior in natural settings, though it is possible that the

accuracy of ASPIRE in the laboratory does not generalize to these settings. Thus, future efforts should focus on demonstrating the applicability of ASPIRE in studying smoking behavior in natural settings.

In summary, this study provides preliminary evidence of ASPIRE's potential to accurately and reliably detect smoking characteristics passively and in real-time. The ability to observe smoking behavior in situ holds great promise in advancing research on the mechanisms that maintain cigarette smoking, measuring behavior change in the context of clinical trials, and the development of novel, real-time interventions for smoking cessation and just-in-time relapse prevention interventions.

#### 5.4 Resolving Ambiguities In Accelerometer Data Due To Location Of Sensor On Wrist In Application to Detection of Smoking Gesture

##### 5.4.1 Abstract

Diseases resulting from prolonged smoking are the most common preventable causes of death in the world today. Automated identification of smoking gestures can help to initiate the appropriate intervention method and prevent relapses in smoking. In previous work, we investigated the success of utilizing accelerometer sensors in smart watches to identify smoking gestures. Our experiments have indicated that identification of smoking gestures is indeed possible with 85%-95% accuracy through the use of Artificial Neural Networks (ANNs). As a follow-up study we present an investigation into the ambiguities in accelerometer data that arise due to the position of the smart watch on a person's wrist. As such, we have developed a method for transforming data to resolve the ambiguities for eight common configurations on the wrist. In this study we have utilized sensor data from the Pebble Time Steel smart watch. Results of our investigation indicate 100% success in

recovery of individual smoking gestures after our developed transformation. Additionally, our results indicate an average increase of 29.6% in detection accuracy when the method is applied to continuous smoking sessions. The methodology created in this work will be integrated into a mobile application for the automated detection of smoking to make it more flexible and robust. Inclusion of this method will greatly increase the accuracy of the ANN when it is faced with varying configurations of the watch.

#### *5.4.2 Introduction*

In the past decade, the subfield of activity recognition has emerged within the field of health informatics. Traditionally this research was conducted with the use of custom engineered devices that were worn across various parts of a subject's body[202, 220], but recently much of these investigations have shifted to utilizing both smart phones[221] and smart watches[188, 222]. Concurrent to this growth in the field of activity recognition was an international effort to warn the population about the dangers of smoking. Although the smoking rate has decreased significantly since then, smoking related diseases are still the most common preventable causes of death in the world today. In addition, tobacco use by both college and high school students has steadily increase as the popularity of product such as e-cigarettes and hookah has risen[177, 178]. On average, smokers relapse four times before successfully quitting[179]. It has been shown that constant support from an individual's community is shown to increase the likelihood of quitting[179]. The existence of an application (housed on a smart phone or watch) that would provide this constant support could greatly increase a person's resoluteness in abstaining from smoking.

The first step in making such an application is the ability to detect when a person is smoking so that the appropriate intervention can be initiated. Identification of a smoking

gesture provides a clearly defined challenge in the field of activity recognition with possibly highly impactful results. Previous works have shown that it is possible to detect smoking gestures using in-house designed wearable devices[173, 174, 191] such as multiple 9-axis inertial measurement units (IMU's), respiration bands and two-lead electrocardiographs worn under the clothes. Whereas these techniques have shown great promise with both high accuracy (95.7-96.9%) and low false positive rates (<1.5%), use of uncommon and relatively expensive devices severely limits mass deployment for daily use. However, in our previous work[192] we have shown that sufficient accuracy (85-95%) and false positive rates (<20%) can be achieved with accelerometer data collected from a common smart watch and a trained Artificial Neural Networks (ANNs).

In the next phase of investigation, a larger study involving several dozen participants was conducted. The particular protocol for this study required participants to collect a specified number of smoking and non-smoking activities while wearing the watch in a pronated configuration on their right wrist. However, some smoking activities, such as smoking and driving, necessitated the participant to use their left hand. In addition, due to user error, there were several participants that wore the watch upside down resulting in deviations from a typical smoking gesture. In these cases, the ANN was unable to detect smoking thus motivating a need for a method to resolve these ambiguities. In this investigation eight common configurations of a smart watch on an individual's wrist and the resulting data transformations were investigated. Prior knowledge of these transformations will allow for a more complete deployment of our detection mechanism on smart watches in the future.

### 5.4.3 Background and Method

#### 5.4.3.1 Data Collection

In this study eight possible configurations of the smart watch on a participant's wrist were explored. A short description and shorthand name designation is given in Table 5.4. Data was collected on the Pebble Time Steal smart watch at a sampling rate of 50Hz using the AccelTool application (<http://mgabor.hu/accel/>). For each of the eight configurations, five instances of a single smoking gesture were collected. This yielded, in total, 40 gestures. A sample from each configuration is shown in Figure. 5.30.

Table 5.4. A short description and shorthand designation are given for each position on the wrist. "Right side up" denotes that the watch is positioned such that text on the screen is not upside down. (L/R) denotes either the left (L) or right (R) wrist.

Designation	Description
(L/R)P1	Right side up in pronated position
(L/R)P2	Upside down in pronated position
(L/R)P3	Right side up in supinated position
(L/R)P4	Upside down in supinated position

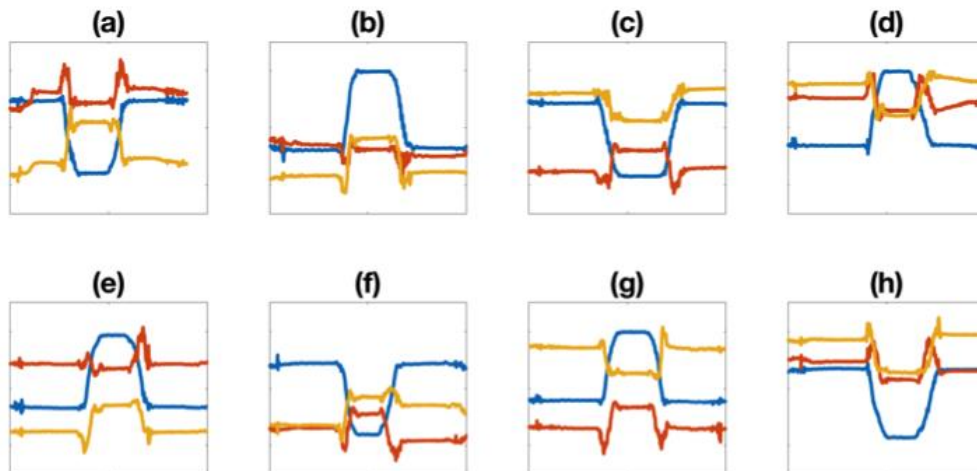


Figure 5.30. XYZ plots for configuration: (a) LP1, (b) LP2, (c) LP3, (d) LP4, (e) RP1, (f) RP2, (g) RP3, and (h) RP4.



In addition to single smoking gestures, a continuous smoking session for each configuration was collected. In this case, a continuous session is defined as an alternating sequence of an individual smoking gesture and a two second resting period. Each continuous session included five individual gestures.

#### 5.4.3.2 Transformation of Data

For this study, the configuration of LP1 was taken as the anchor position since data from this configuration was used to train the neural network from the original work[12]. Although conceivably different ANNs can be trained for each configuration, this is not an ideal approach given the limited resources in mobile devices (especially so for smart watches) and acquisition of additional data for training of each network. The more optimal approach is to provide a transformation layer that will map each set of data to the common frame of the LP1. Given the nature of different configurations of the watch, such a transformation could be theorized based on two components. The first component of this transformation should correct for the altered orientation of the watch in other configurations. Rotational operators can implement this first transformation. Here we utilized Eulerian rotations of the Cartesian sensor data to relate all configurations to the anchor frame. Equation (5.4.1) shows the Euler transformation that permits full rotation of the space using three consecutive rotations about z, y, z axes. Rotation about each of the axes is denoted by  $\alpha$ ,  $\beta$ ,  $\gamma$  where  $\gamma$  signifies the first rotation (note that rotational transformations are not commutative). Table 5.5 describes the values of  $\alpha$ ,  $\beta$ ,  $\gamma$  that correspond to the reorientation component of each configuration with respect to the anchor configuration under the ideal anatomical relationship (perfect symmetry of posture and etc.).

The second component of a complete transformation should account for the alternate path that is traversed from rest-to-mouth during a smoking gesture (or a puff). This transformation is only required when the watch is placed on the opposite arm to the reference point (LP1 in this work). This transformation in chirality of the sensor space can be obtained by comparing the similar gestures from LP1 (our reference configuration) and RP4 since the orientation of the watch is the same in both configurations. Therefore, any alternation in the sensor data must be due to the change in chirality of the traversed path of the smartwatch. The resultant chirality inversion transformation is shown in Equation (5.4.2) and the complete transformation of the sensor data is shown in Equation (5.4.3). Figure 5.31 shows the individual gesture samples in Figure 5.30 after proper transformation using the information shown in Table 5.5 and Equation (5.4.3).

$$\begin{pmatrix} \cos(\alpha)\cos(\beta)\cos(\gamma) - \sin(\alpha)\sin(\gamma) & -\cos(\alpha)\cos(\beta)\sin(\gamma) - \sin(\alpha)\cos(\gamma) & \cos(\alpha)\sin(\beta) \\ \sin(\alpha)\cos(\beta)\cos(\gamma) + \cos(\alpha)\sin(\gamma) & -\sin(\alpha)\cos(\beta)\sin(\gamma) + \cos(\alpha)\cos(\gamma) & \sin(\alpha)\sin(\beta) \\ -\sin(\beta)\cos(\gamma) & \sin(\beta)\sin(\gamma) & \cos(\beta) \end{pmatrix} \quad (5.4.1)$$

$$C_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \quad (5.4.2)$$

$$S' = C_i \cdot R(\alpha, \beta, \gamma) \quad (5.4.3)$$

Table 5.5. The values for  $\alpha$ ,  $\beta$  and  $\gamma$  are given for each of the configurations.

Position	$\alpha_{z''}$	$\beta_{y'}$	$\gamma_z$
LP1	0°	0°	0°
LP2	0°	0°	180°
LP3	0°	180°	180°
LP4	0°	180°	0°

RP1	0°	180°	0°
RP2	0°	180°	180°
RP3	0°	0°	180°
RP4	0°	0°	0°

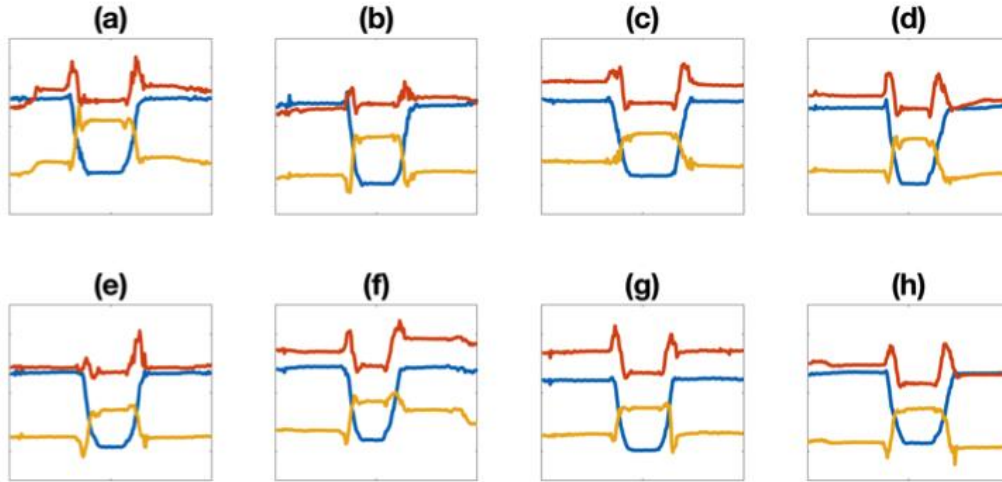


Figure 5.31. XYZ plots for rotated positions: (a) LP1, (b) LP2, (c) LP3, (d) LP4, (e) RP1, (f) RP2, (g) RP3, and (h) RP4.

Visual inspection can confirm that the resulting samples closely resemble the anchor configuration.

#### 5.4.3.3 Evaluation on Previously Trained ANN

The neural network toolbox in Matlab (version R2016a) was utilized during this phase of the study. In previous work, a series of ANNs were trained using individual smoking gestures and a variety of non-smoking gestures with the user wearing the watch in LP1. In this study each individual dimension of the accelerometer data was isolated and explored. Neural networks were created for the X, Y, and Z dimension both separately (X-ANN, Y-ANN, Z-ANN) and combined (XYZ-ANN). This work concluded that the X

dimension provided the most information for detection of an individual smoking gesture. However, when the X-ANN was presented with more complex non-smoking gestures such as eating and drinking, it produced a high number of false positives. In the same study, the XYZ-ANN maintained a high accuracy alongside the X-ANN while also maintained a low false positive rate. Therefore, it was concluded that inclusion of the Y and Z dimension was necessary to achieve consistent high accuracy (>75%) and an acceptable false positive rate (<20%). For this reason, the current study focused on the XYZ-ANN.

#### 5.4.3.4 Evaluation of The Proposed Transformations

*5.4.3.4.1 Individual Gesture Detection Evaluation*—The original data as well as the transformed data were classified using the XYZ-ANN. If an output of 0.75 or greater was observed from the network then the signal was classified as a smoking gesture. Results of this exercise are summarized in Section 5.4.4.1.

*5.4.3.4.1 Continuous Session Detection Evaluation*—The first step in evaluating a continuous session was to identify the positions in the session where the smoking gestures occurred. There was no automated method of accomplishing this task and it was therefore completed manually. Since the annotation of the start and end of a smoking gesture is highly subjective, evaluation of a continuous smoking session is difficult to quantify. In this study the start and end of a smoking gesture is defined as the area in which the X dimension exhibits a dip (this dip can be seen clearly in Fig. 5.4.1(a)). Any output within these dip regions greater than 0.75 was considered a true positive (anything <0.75 was classified as a false negative). Any regions outside of these ranges were considered non-smoking gestures. Outputs in these regions less than 0.75 was considered a true negative (anything >0.75 was classified a false positive). The accuracy was then calculated using

Eq. (5.4.4) where TP denotes the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5.4.4)$$

As with the individual gestures, the original and transformed data were classified using the XYZ-ANN. The results are summarized in Section 5.4.4.2.

#### 5.4.4 Results and Discussion

The results of simulating the XYZ-ANN on the test data are reported in the following sections. In each section a short discussion of the results is also included.

##### 5.4.4.1 Detection of Individual Smoking Gesture

For each of the eight configurations, the number of correctly detected smoking gestures was recorded before and after transformation of the data. Fig. 5.32 shows the resulting counts of true positives.

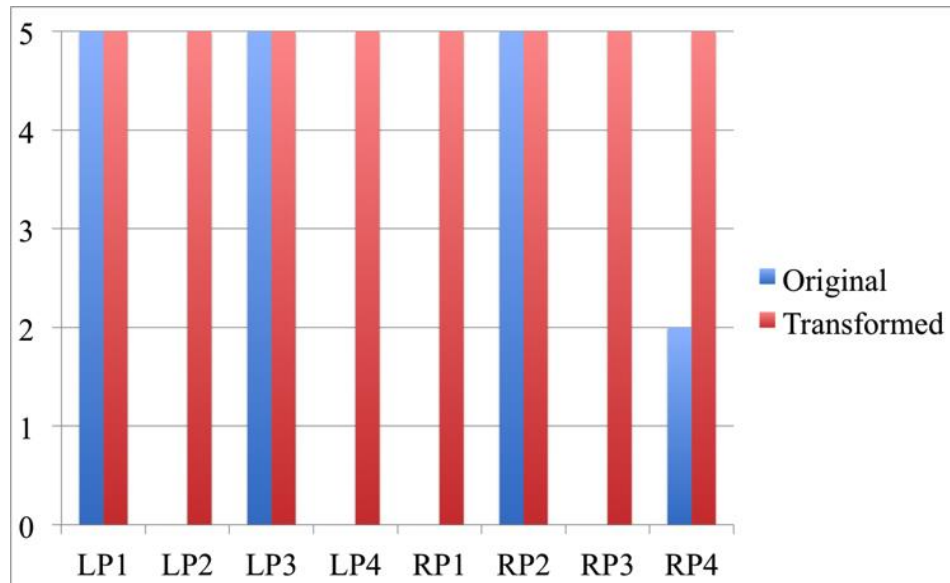


Figure 5.32. The number of correctly identified smoking gestures before transformation is denoted by the blue bars. The signals after transformation are represented by the red bars.

As seen in Figure 5.32, for each of the eight configurations, transformation of the data resulted in all five of the gestures being identified as smoking gestures. It is worth noting that in the cases of LP3, RP2, and RP4 some or all of the five gestures were correctly identified as smoking before the transformation. One possible explanation is the fact that in these three conformations the X dimension remains fixed (see Table 5.5). This is consistent with the previous observation that the X dimension is the most informational in application to detection of smoking and since it is not changed in these configurations, smoking gestures are still detected before the transformation.

#### 5.4.4.2 Detection of Smoking in Continuous Sessions

Table 5.6 shows the accuracies in detection of each individual smoking gesture within each continuous session before and after transformation. In addition, a percentage change is shown highlighting the change in performance before and after transformation.

In all cases the percentage change is positive, signifying the effectiveness of the developed transformations. In addition, the accuracies are comparable to the accuracies achieved in previous work with the XYZ-ANN using only LP1 data. An example of the improvement in detection is shown in Figure 5.33. Recall that a smoking gesture was defined as the area between the start and end of a dip in the X dimension (in blue). A clear refinement can be observed between pane (a) and pane (b) in the areas where the XYZ-

ANN detects smoking (denoted by a bump in the black line). In pane (b) more of each smoking gesture is encapsulated by the spikes in black.

Table 5.6. Accuracies in detection of individual smoking gestures inside of continuous sessions before and after transformation are reported.

Configuration	Before	After	% Change
LP1	73.23	–	–
LP2	26.98	78.45	+51.47
LP3	66.89	78.69	+11.80
LP4	34.03	76.27	+42.24
RP1	31.95	67.04	+35.09
RP2	70.43	71.78	+1.35
RP3	23.54	76.20	+52.66
RP4	63.95	76.33	+12.38

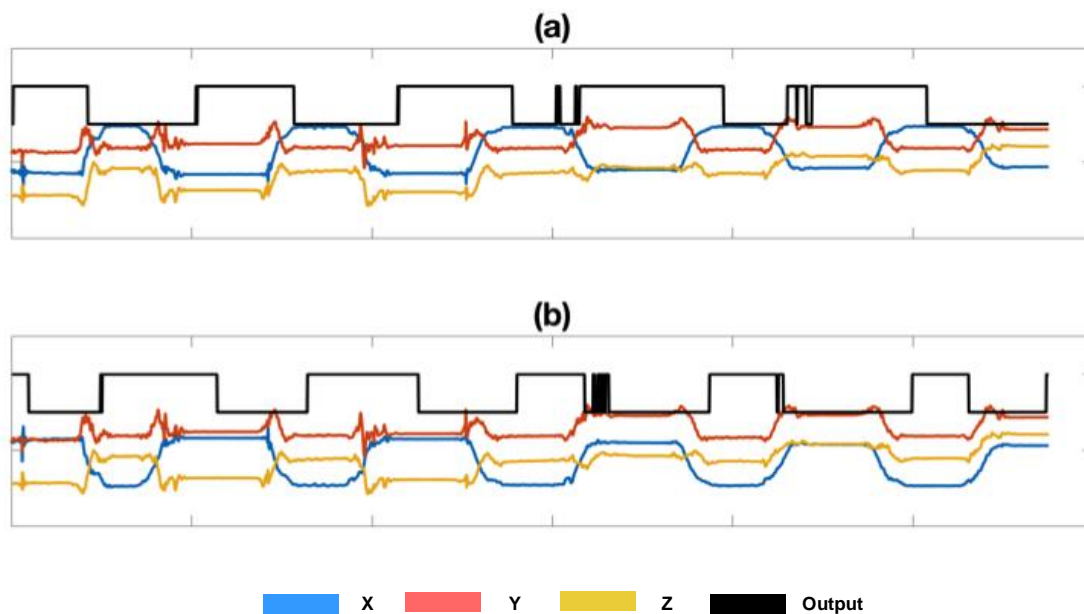


Figure 5.33. Continuous session with watch in RP1 configuration (a) before and (b) after transformation is depicted. Output of the XYZ-ANN is shown in black.

#### 5.4.5 Conclusion

Based on the results from Figure 5.33 and Table 5.6 it can be concluded that our proposed methodology is a suitable approach for resolving the ambiguities that arise from varying the configuration of a watch worn on the wrist. In addition, this study has produced further evidence that the X dimension of the accelerometer data is the most informational in detection of smoking. Evidence for this conclusion comes from the results of LP3, RP2, and RP4. In each of these configurations the X dimension is not rotated in the original data. These configurations show the highest degree of success in individual detection before transformation and show only modest gains after transformation in detection of continuous sessions. This confirms that inclusion of the Y and Z dimensions do in fact increase the accuracy, but are not as important as X.

Most left-handed individuals wear watches on the right wrist instead of the left thus, arguably, the most common configuration switch will be that of LP1 to RP1. Not only did the transformation of RP1 produce 100% accuracy in individual gesture detection it caused a 35.09% increase in accuracy in detection of continuous gestures. In a future smoking detection app, the watch configuration preference can be entered by the user. The app can then leverage this setting to do the correct transformations to ensure detection accuracy. It is possible to automate this process in the future by including a preliminary calibration step in which the watch will automatically detect the configuration of the watch.



## Chapter 6: Summarization of Major Results for the Automated Detection of Smoking

### 6.1 Puff Level Binary Detection of Smoking

The first objective of this part of my research was to create a model that performs a binary detection of a smoking puff. A full description of the work completed can be found in section 5.1. The following is a summary of the major achievements. Binary classification (“smoking” vs. “non-smoking”) was completed using a traditional ANN. The network architecture was a 300 neuron input layer, followed by a 10 neuron hidden layer with a 1 neuron output. The network was initially trained with smoking data collected in a laboratory setting. Later, data from real smokers was also added to the training set and the model was retrained and tested. For both iterations of the model, an accuracy of nearly 90% was achieved. In an effort to reduce the total number of neurons in the input layer several factors were considered and explored. The first of which was whether all axes of the accelerometer needed to be used to maintain accuracy. In our initial studies we found that the x-axis was the most informative when considering single gestures, however, when the network was deployed to extract single gestures out of continuous smoking session, we found that using data from all three axes yielded better results. The second factor that was investigated was the sampling rate. Our initial studies used a sampling rate of 100Hz however we found that this rate can be dropped to as low as 25Hz while still maintaining over 87% accuracy.

This work marked the first study to be completed in the arena of automated smoking detection using smartwatch technology. After the completion of this work, a patent was applied for and awarded (Patent number: US 10551935).

As promising as these results were, there were still one clear bias in detection of the smoking puffs when the model was applied to larger, more complex datasets. We found that the binary classifier detected the first edge of smoking very accurately, however accuracy dropped off towards the end of each puff. However, this bias was addressed in future works reported in this work.

## 6.2 Session Level Detection of Smoking

The second objective of this part of my research was to extend the binary detection of smoking and apply it to the detection of smoking sessions (or full a cigarette). A full description of the work completed can be found in section 5.2. The following is a summary of the major achievements. As stated previously the first challenge of this objective was that a smoking session was not analytically defined at the time the work started. Therefore, utilizing both subject matter experts and empirical data collected from real smokers, a model of a smoking session was developed. Session level detection was accomplished by utilizing the results from the puff level detection in combination with a rule-based AI. The study performed to validate the model included the recruitment of ten heavy smokers to wear a smartwatch that housed our in-house sensor recording app for several days to record data. The participants also utilized a Google Form as a mechanism of marking the start and end times of each smoking session throughout the day. The true positive rate reported from this model was over 89% once misclassified sessions resulting from self-reporting errors were excluded. The false positive rate was 2.1%. Again, the results from this work mark a

first in several aspects. The analytical model created to describe a smoking event (or cigarette) is the first of its kind and is currently being employed by several smoking cessation researchers throughout the state.

### 6.3 Automated Characterization of Smoking Topography

The third objective of this part of my research was to redevelop the binary model presented in section 5.1 to not just detect, but to also characterize several topographies of smoking. A full description of the work completed can be found in section 5.3. The following is a summary of the major achievements. The first step in redevelopment was to reannotate the training set from the previous studies. In those studies, smoking gestures were separated into full puffs. However, in order to overcome the bias of the binary classifier and characterize interesting features of smoking such as duration and time between puffs, the data needed to be dissected further. The data was reformulated into “mini” gestures (“hand-to-lip”, “hand-on-lip”, “hand-off-lip” and “non-smoking”). After successful reannotation of the data, a conventional neural network was employed that had an output size of four (one for each “mini” gesture). The output of this network was then fed to a rule-based AI that modelled the smoking behavior as a state transition model. As a prerequisite to deploying a full neural network on the problem, proof of concept study was performed. In this work, smoking data collected from a smartwatch was compared to that of data collected from the CReSS smoking device. Of particular note, the correlation between the puff duration reported by CReSS and our model was 0.79 and the correlation of the time between puffs (inter-puff interval) was 0.73. This work has the potential to revolutionize the field of tobacco related research. Up to this point, to collection of this

level of data was only possible via laboratory settings. Smoking in this setting is seen as unnatural to smokers and therefore causes biases in the data collected.

#### 6.4 Resolving Ambiguities in Accelerometer Data Based on the Position of Smartwatch on Wrist

The final objective of this part of my research was to develop a method for resolving the ambiguities in accelerometer data due to position on the wrist. A full description of the work completed can be found in section 5.4. The following is a summary of the major achievements. This was a very important step in processing the data for use in an ANN as each position of the watch on the wrist results in different signals. We defined eight distinct configurations of watch placement on the wrist. Instead of insisting on a particular configuration to be used across all users, a transformation technique was created to rotate data coming from all other configurations to a common reference frame for use in the ANN. Our results for this work indicate that for single puff detection our method had 100% accuracy in detecting transformed puffs. Results from detection of smoking puffs within a continuous smoking also showed very high success. The baseline accuracy was 73%. After transformation, all but two of the eight positions recorded accuracies of 73% or greater.

#### 6.5 Suggested Future Work

My suggested future work for these objectives would focus around enhanced model development. Although several other types of predictive models were considered at the beginning of development, none were investigated fully. The models used in this work were conventional single layer neural networks. With the recent resurgence of deep-learning and more complicated neural networks, it should be noted that several of these models could hold the potential to increase the accuracy of the detection models in more

efficient ways. In addition, although this work has been centered on the detection of smoking, similar detection models could be created for other gestures (such as overeating and alcohol consumption) using the same methods described throughout the work.

## Chapter 7: Conclusion

This work was divided into two main parts that spanned the broad field of computational biology and bioinformatics. In part 1, it was shown that substantial improvement to the software package REDCRAFT, creation of the PDBMine database and creation of an analytic model for the characterization of protein dynamics has resulted in an increased capability for protein structure/dynamic calculation. Of note, this work presents a novel database for mining data from proteomic data as well as the first known theoretical model of discrete state dynamics from RDCs. In part 2 of this work, several novel methods for detection of smoking in real world settings was presented. These models showed substantial success in detection of puff-level and session-level detection of smoking events. In addition, a method for quantification of smoking behavior was also presented. This method showed significant similarity to the devices currently used in the field for data collection. Lastly, a method for the rotational translation of accelerometer data to resolve ambiguities due to the position of the watch on the wrist. These results represent the first published attempt to create a method for automated detection of smoking using smartwatch technology.

## References

- [1] H. M. Berman *et al.*, "The protein structure initiative structural genomics knowledgebase.," *Nucleic acids research*, vol. 37, pp. D365-8, 2009.
- [2] H. P. J. Buermans and J. T. den Dunnen, "Next generation sequencing technology: Advances and applications.," *Biochimica et biophysica acta*, vol. 1842, pp. 1932-1941, 2014.
- [3] J. R. Schnell, H. J. Dyson, and P. E. Wright, "Structure, dynamics, and catalytic function of dihydrofolate reductase.," *Annual review of biophysics and biomolecular structure*, vol. 33, pp. 119-40, 2004.
- [4] P. M. Bowers, C. E. Strauss, and D. Baker, "De novo protein structure determination using sparse NMR data.," *Journal of biomolecular NMR*, vol. 18, pp. 311-8, 2000.
- [5] J. R. Tolman, H. M. Al-Hashimi, L. E. Kay, and J. H. Prestegard, "Structural and dynamic analysis of residual dipolar coupling data for proteins.," *Journal of the American Chemical Society*, vol. 123, pp. 1416-24, 2001.
- [6] A. Bax, "Weak alignment offers new NMR opportunities to study protein structure and dynamics.," *Protein science : a publication of the Protein Society*, vol. 12, pp. 1-16, 2003, doi: 10.1110/ps.0233303.
- [7] M. Blackledge, "Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 46, pp. 23-61, 2005, doi: 10.1016/j.pnmrs.2004.11.002.
- [8] S. Lee, M. F. Mesleh, and S. J. Opella, "Structure and dynamics of a membrane protein in micelles from three solution NMR experiments," *J Biomol NMR*, vol. 26, pp. 327-334, 2003.
- [9] H. M. Berman *et al.*, "The Protein Data Bank," *Acta Crystallogr D Biol Crystallogr*, vol. 58, no. Pt 6 No 1, pp. 899-907, Jun 2002. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/12037327>.
- [10] G. Bouvignies, S. Meier, S. Grzesiek, and M. Blackledge, "Ultrahigh-resolution backbone structure of perdeuterated protein GB1 using residual dipolar couplings

- from two alignment media," *Angew Chem Int Ed Engl*, vol. 45, pp. 8166-8169, 2006, doi: 10.1002/anie.200603627.
- [11] A. Yershova, C. Tripathy, P. Zhou, and B. R. Donald, "Algorithms and Analytic Solutions Using Sparse Residual Dipolar Couplings for High-Resolution Automated Protein Backbone Structure Determination by NMR," ed, 2010, pp. 355-372.
- [12] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, pp. 187-217, 1983, doi: 10.1002/jcc.540040211.
- [13] D. A. Case *et al.*, "The Amber biomolecular simulation programs," *Journal Of Computational Chemistry*, vol. 26, pp. 1668-1688, 2005.
- [14] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," *Journal of Chemical Theory and Computation*, vol. 4, pp. 435-447, 2008, doi: 10.1021/ct700301q.
- [15] J. C. Phillips *et al.*, "Scalable molecular dynamics with NAMD," *Journal Of Computational Chemistry*, vol. 26, pp. 1781-1802, 2005.
- [16] M. Bryson, F. Tian, J. H. Prestegard, and H. Valafar, "REDCRAFT: a tool for simultaneous characterization of protein backbone structure and motion from RDC data.," *Journal of Magnetic Resonance*, vol. 191, pp. 322-34, 2008, doi: 10.1016/j.jmr.2008.01.007.
- [17] H. Valafar, M. Simin, and S. Irausquin, "A Review of REDCRAFT," in *Annual Reports on NMR Spectroscopy* vol. 76, ed, 2012, pp. 23-66.
- [18] F. Tian, H. Valafar, and J. H. Prestegard, "A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones.," *Journal of the American Chemical Society*, vol. 123, pp. 11791-6, 2001.
- [19] J. H. Prestegard, K. L. Mayer, H. Valafar, and G. C. Benison, "Determination of protein backbone structures from residual dipolar couplings.," *Methods in enzymology*, vol. 394, pp. 175-209, 2005, doi: 10.1016/S0076-6879(05)94007-X.
- [20] P. Shealy, M. Simin, S. H. Park, S. J. Opella, and H. Valafar, "Simultaneous structure and dynamics of a membrane protein using REDCRAFT: membrane-bound form of Pf1 coat protein.," *Journal of Magnetic Resonance*, vol. 207, pp. 8-16, 2010, doi: 10.1016/j.jmr.2010.07.016.



- [21] Y. Shen, F. Delaglio, G. Cornilescu, and A. Bax, "TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts," *Journal of Biomolecular NMR*, vol. 44, pp. 213-223, 2009, doi: 10.1007/s10858-009-9333-z.
- [22] A. Gutmanas *et al.*, "NMR Exchange Format: a unified and open standard for representation of NMR restraint data.," *Nature structural & molecular biology*, vol. 22, pp. 433-4, 2015, doi: 10.1038/nsmb.3041.
- [23] P. Dosset, J. C. Hus, D. Marion, and M. Blackledge, "A novel interactive tool for rigid-body modeling of multi-domain macromolecules using residual dipolar couplings.," *Journal of biomolecular NMR*, vol. 20, pp. 223-31, 2001.
- [24] M. a. Martí-Renom, a. C. Stuart, a. Fiser, R. Sánchez, F. Melo, and a. Sali, "Comparative protein structure modeling of genes and genomes.," *Annual review of biophysics and biomolecular structure*, vol. 29, pp. 291-325, 2000, doi: 10.1146/annurev.biophys.29.1.291.
- [25] N. Rodriguez, D. Vinal, J. Rodriguez-Cobos, J. De Castro, and G. Dominguez, "Genomic profiling in oncology clinical practice," *Clin Transl Oncol*, Jan 24 2020, doi: 10.1007/s12094-020-02296-9.
- [26] S. Morganti *et al.*, "Complexity of genome sequencing and reporting: Next generation sequencing (NGS) technologies and implementation of precision medicine in real life," *Crit Rev Oncol Hematol*, vol. 133, pp. 171-182, Jan 2019, doi: 10.1016/j.critrevonc.2018.11.008.
- [27] S. Morganti, P. Tarantino, E. Ferraro, P. D'Amico, B. A. Duso, and G. Curigliano, "Next Generation Sequencing (NGS): A Revolutionary Technology in Pharmacogenomics and Personalized Medicine in Cancer," *Adv Exp Med Biol*, vol. 1168, pp. 9-30, 2019, doi: 10.1007/978-3-030-24100-1\_2.
- [28] R. Kamps *et al.*, "Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification," *International Journal of Molecular Sciences*, vol. 18, p. 308, 2017, doi: 10.3390/ijms18020308.
- [29] C. Manzoni *et al.*, "Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences," *Brief Bioinform*, vol. 19, no. 2, pp. 286-302, Mar 1 2018, doi: 10.1093/bib/bbw114.
- [30] J. Martin and A. Sawyer, "Elucidating the structure of membrane proteins," (in English), *Biotechniques*, vol. 66, no. 4, pp. 167-170, Apr 2019, doi: 10.2144/btn-2019-0030.
- [31] D. S. Goodsell *et al.*, "RCSB Protein Data Bank: Enabling biomedical research and drug discovery," *Protein Sci*, vol. 29, no. 1, pp. 52-65, Jan 2020, doi: 10.1002/pro.3730.

- [32] A. Pandey, K. Shin, R. E. Patterson, X. Q. Liu, and J. K. Rainey, "Current strategies for protein production and purification enabling membrane protein structural biology," *Biochem Cell Biol*, vol. 94, no. 6, pp. 507-527, Dec 2016, doi: 10.1139/bcb-2015-0143.
- [33] D. Hardy, R. M. Bill, A. Jawhari, and A. J. Rothnie, "Overcoming bottlenecks in the membrane protein structural biology pipeline," *Biochem Soc Trans*, vol. 44, no. 3, pp. 838-44, Jun 15 2016, doi: 10.1042/BST20160049.
- [34] S. J. Opella, "Structure determination of membrane proteins by NMR spectroscopy," *Abstracts of Papers of the American Chemical Society*, vol. 214, pp. 179-PHYS, 1997.
- [35] S. J. Opella and F. M. Marassi, "Applications of NMR to membrane proteins," *Arch Biochem Biophys*, vol. 628, pp. 92-101, Aug 15 2017, doi: 10.1016/j.abb.2017.05.011.
- [36] Y. Tian, C. D. Schwieters, S. J. Opella, and F. M. Marassi, "High quality NMR structures: a new force field with implicit water and membrane solvation for Xplor-NIH," *J Biomol NMR*, vol. 67, no. 1, pp. 35-49, Jan 2017, doi: 10.1007/s10858-016-0082-5.
- [37] S. J. Opella, "Solid-state NMR and membrane proteins," *J Magn Reson*, vol. 253, pp. 129-37, Apr 2015, doi: 10.1016/j.jmr.2014.11.015.
- [38] S. Bansal, X. Miao, M. W. W. Adams, J. H. Prestegard, and H. Valafar, "Rapid classification of protein structure models using unassigned backbone RDCs and probability density profile analysis (PDPA)." *Journal of Magnetic Resonance*, vol. 192, pp. 60-8, 2008, doi: 10.1016/j.jmr.2008.01.014.
- [39] J. R. Tolman, "Dipolar couplings as a probe of molecular dynamics and structure in solution," *Current Opinion in Structural Biology*, vol. 11, pp. 532-539, 2001.
- [40] K. Lindorff-Larsen, R. B. Best, M. a. Depristo, C. M. Dobson, and M. Vendruscolo, "Simultaneous determination of protein structure and dynamics.," *Nature*, vol. 433, pp. 128-32, 2005, doi: 10.1038/nature03199.
- [41] G. Bouvignies, P. R. L. Markwick, and M. Blackledge, "Simultaneous definition of high resolution protein structure and backbone conformational dynamics using NMR residual dipolar couplings.," *Chemphyschem : a European journal of chemical physics and physical chemistry*, vol. 8, pp. 1901-1909, 2007, doi: 10.1002/cphc.200700353.
- [42] C. A. Cole, R. Mukhopadhyay, H. Omar, M. Hennig, and H. Valafar, "Structure Calculation and Reconstruction of Discrete-State Dynamics from Residual Dipolar

- Couplings.," *Journal of chemical theory and computation*, vol. 12, pp. 1408-22, 2016, doi: 10.1021/acs.jctc.5b01091.
- [43] S. Olsson, D. Ekonomiuk, J. Sgrignani, and A. Cavalli, "Molecular Dynamics of Biomolecules through Direct Analysis of Dipolar Couplings.," *Journal of the American Chemical Society*, vol. 137, pp. 6270-6278, 2015.
- [44] M. Rinaldelli, E. Ravera, V. Calderone, G. Parigi, G. N. Murshudov, and C. Luchinat, "Simultaneous use of solution NMR and X-ray data in REFMAC5 for joint refinement/detection of structural differences.," *Acta crystallographica. Section D, Biological crystallography*, vol. 70, pp. 958-67, 2014, doi: 10.1107/S1399004713034160.
- [45] R. W. Montalvao, A. D. Simone, and M. Vendruscolo, "Determination of structural fluctuations of proteins from structure-based calculations of residual dipolar couplings," *J Biomol NMR*, vol. 53, pp. 281-292, 2012, doi: 10.1007/s10858-012-9644-3.
- [46] M. Simin, S. Irausquin, C. A. Cole, and H. Valafar, "Improvements to REDCRAFT: a software tool for simultaneous characterization of protein backbone structure and dynamics from residual dipolar couplings.," *Journal of biomolecular NMR*, vol. 60, pp. 241-64, 2014, doi: 10.1007/s10858-014-9871-x.
- [47] J.-C. Hus, D. Marion, and M. Blackledge, "Determination of protein backbone structure using only residual dipolar couplings.," *Journal of the American Chemical Society*, vol. 123, pp. 1541-2, 2001.
- [48] F. Delaglio, G. Kontaxis, and A. Bax, "Protein Structure Determination Using Molecular Fragment Replacement and NMR Dipolar Couplings," *Journal of the American Chemical Society*, vol. 122, pp. 2142-2143, 2000, doi: 10.1021/ja993603n.
- [49] M. Andrec, P. Du, and R. M. Levy, "Protein backbone structure determination using only residual dipolar couplings from one ordering medium," *Journal of Biomolecular NMR*, vol. 21, pp. 335-347, 2001, doi: 10.1023/A:1013334513610.
- [50] S. Yang and H. M. Al-Hashimi, "Unveiling Inherent Degeneracies in Determining Population-Weighted Ensembles of Interdomain Orientational Distributions Using NMR Residual Dipolar Couplings: Application to RNA Helix Junction Helix Motifs.," *The journal of physical chemistry. B*, vol. 119, pp. 9614-26, 2015.
- [51] K. Chen and N. Tjandra, "The use of residual dipolar coupling in studying proteins by NMR.," *Topics in current chemistry*, vol. 326, pp. 47-67, 2012, doi: 10.1007/128\_2011\_215.

- [52] J. Zeng *et al.*, "High-resolution protein structure determination starting with a global fold calculated from exact solutions to the RDC equations.," *Journal of biomolecular NMR*, vol. 45, pp. 265-81, 2009.
- [53] C. SCHWIETERS, J. KUSZEWSKI, and G. MARIUSCLORE, "Using Xplor–NIH for NMR molecular structure determination," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 48, pp. 47-62, 2006, doi: 10.1016/j.pnmrs.2005.10.001.
- [54] X. Wang, B. Tash, J. M. Flanagan, and F. Tian, "RDC derived protein backbone resonance assignment using fragment assembly.," *Journal of biomolecular NMR*, vol. 49, pp. 85-98, 2011, doi: 10.1007/s10858-010-9467-z.
- [55] C. A. Cole, C. Ott, D. Valdes, and H. Valafar, "PDBMine: A Reformulation of the Protein Data Bank to Facilitate Structural Data Mining," presented at the IEEE Annual Conf. on Computational Science & Computational Intelligence (CSCI), Las Vegas, NV, December 5th-7th, 2019, 2019.
- [56] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, "Stereochemistry of polypeptide chain configurations," *Journal of Molecular Biology*, vol. 7, pp. 95-99, 1963, doi: 10.1016/S0022-2836(63)80023-6.
- [57] A. Saupe and G. Englert, "High-Resolution Nuclear Magnetic Resonance Spectra of Orientated Molecules," *Physical Review Letters*, vol. 11, pp. 462-464, 1963, doi: 10.1103/PhysRevLett.11.462.
- [58] H. Lee, P. K. S. T, G. T. Montelione, and J. H. Prestegard, "Alignment Media Preparation," 2013.
- [59] A. Bax, G. Kontaxis, and N. Tjandra, "Dipolar couplings in macromolecular structure determination.," *Methods in enzymology*, vol. 339, pp. 127-74, 2001.
- [60] K. Chen and N. Tjandra, "Top-down approach in protein RDC data analysis: de novo estimation of the alignment tensor.," *Journal of biomolecular NMR*, vol. 38, pp. 303-13, 2007, doi: 10.1007/s10858-007-9168-4.
- [61] M. Clemencic and P. Mato, "A CMake-based build and configuration framework," (in English), *J Phys Conf Ser*, vol. 396, 2012, doi: Artn 052021
- [62] H. M. Berman *et al.*, "The Protein Data Bank," *Nucleic Acids Res*, vol. 28, no. 1, pp. 235-42, Jan 1 2000, doi: 10.1093/nar/28.1.235.
- [63] M. Andrec, P. C. Du, and R. M. Levy, "Protein backbone structure determination using only residual dipolar couplings from one ordering medium.," *Journal of biomolecular NMR*, vol. 21, pp. 335-47, 2001.

- [64] G. M. Clore and C. D. Schwieters, "How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation?," *Journal of the American Chemical Society*, vol. 126, pp. 2923-38, 2004.
- [65] D. Long and R. Brüschweiler, "In silico elucidation of the recognition dynamics of ubiquitin.," *PLoS computational biology*, vol. 7, p. e1002035, 2011, doi: 10.1371/journal.pcbi.1002035.
- [66] C. Tripathy, J. Zeng, P. Zhou, and B. R. Donald, "Protein loop closure using orientational restraints from NMR data.," *Proteins*, vol. 80, pp. 433-453, 2012.
- [67] B. L. Eggimann, V. V. Vostrikov, G. Veglia, and J. I. Siepmann, "Modeling helical proteins using residual dipolar couplings, sparse long-range distance constraints and a simple residue-based force field," *Theor Chem Acc*, vol. 132, no. 10, p. 1388, Oct 1 2013, doi: 10.1007/s00214-013-1388-y.
- [68] E. de Alba and N. Tjandra, "Residual dipolar couplings in protein structure determination," *Methods Mol Biol*, vol. 278, pp. 89-106, 2004, doi: 10.1385/1-59259-809-9:089.
- [69] R. Mukhopadhyay, X. Miao, P. Shealy, and H. Valafar, "Efficient and accurate estimation of relative order tensors from lambda-maps.," *Journal of magnetic resonance (San Diego, Calif. : 1997)*, vol. 198, pp. 236-47, 2009, doi: 10.1016/j.jmr.2009.02.014.
- [70] C. Schmidt, S. J. Irausquin, and H. Valafar, "Advances in the REDCAT software package.," *BMC bioinformatics*, vol. 14, p. 302, 2013, doi: 10.1186/1471-2105-14-302.
- [71] P. Shealy and H. Valafar, "Multiple structure alignment with msTALI.," *BMC bioinformatics*, vol. 13, p. 105, 2012, doi: 10.1186/1471-2105-13-105.
- [72] S. K. Burley *et al.*, "Structural genomics: beyond the human genome project.," *Nature genetics*, vol. 23, pp. 151-7, 1999, doi: 10.1038/13783.
- [73] M. W. W. Adams *et al.*, "The Southeast Collaboratory for Structural Genomics: A high-throughput gene to structure factory," *Accounts of Chemical Research*, vol. 36, pp. 191-198, 2003.
- [74] M. A. Jensen, V. Ferretti, R. L. Grossman, and L. M. Staudt, "The NCI Genomic Data Commons as an engine for precision medicine," *Blood*, vol. 130, no. 4, pp. 453-459, Jul 27 2017, doi: 10.1182/blood-2017-03-735654.
- [75] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25-29, 2000, doi: 10.1038/75556.

- [76] C. B. Devaun McFarland, Benjamin Mueller, Homayoun Valafar, Columbia, SC USA. Applying MSTALI to active site identification studies.
- [77] S. Dayalan, N. D. Gooneratne, S. Bevinakoppa, and H. Schroder, "Dihedral angle and secondary structure database of short amino acid fragments," *Bioinformation*, vol. 1, no. 3, pp. 78-80, Jan 1 2006, doi: 10.6026/97320630001078.
- [78] D. S. Berkholz, P. B. Krenesky, J. R. Davidson, and P. A. Karplus, "Protein Geometry Database: a flexible engine to explore backbone conformations and their relationships to covalent geometry," *Nucleic Acids Res*, vol. 38, no. Database issue, pp. D320-5, Jan 2010, doi: 10.1093/nar/gkp1013.
- [79] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577-2637, 1983, doi: 10.1002/bip.360221211.
- [80] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC Bioinformatics*, vol. 9, p. 40, 2008.
- [81] D. E. Kim, D. Chivian, and D. Baker, "Protein structure prediction and analysis using the Robetta server.," *Nucleic acids research*, vol. 32, pp. W526-31, 2004, doi: 10.1093/nar/gkh468.
- [82] A. Fahim, R. Mukhopadhyay, R. Yandle, J. H. Prestegard, and H. Valafar, "Protein Structure Validation and Identification from Unassigned Residual Dipolar Coupling Data Using 2D-PDPA.," *Molecules (Basel, Switzerland)*, vol. 18, pp. 10162-88, 2013, doi: 10.3390/molecules180910162.
- [83] S. C. Lovell *et al.*, "Structure validation by C $\alpha$  geometry: phi,psi and C $\beta$  deviation.," *Proteins*, vol. 50, pp. 437-50, 2003.
- [84] T. U. Consortium, "The universal protein resource (UniProt)," *Nucleic Acids Res*, vol. 36, pp. D190-5, 2008.
- [85] B. K. Ho and R. Brasseur, "The Ramachandran plots of glycine and pre-proline," *BMC Struct Biol*, vol. 5, p. 14, Aug 16 2005, doi: 10.1186/1472-6807-5-14.
- [86] C. A. Cole, D. Ishimaru, M. Hennig, and H. Valafar. An Investigation of Minimum Data Requirement for Successful Structure Determination of Pf2048.1 with REDCRAFT.
- [87] C. A. C. Hanin Omar, Mirko Hennig, Homayoun Valafar, Columbia, SC USA. Characterization of discrete state dynamics from residual dipolar couplings using REDCRAFT.

- [88] E. Timko, P. Shealy, M. Bryson, and H. Valafar. Minimum Data Requirements and Supplemental Angle Constraints for Protein Structure Prediction with REDCRAFT.
- [89] C. A. Cole, C. Parks, J. Rachele, and H. Valafar, "Improvements of the REDCRAFT Software Package," presented at the Proceedings of the International Conference on Bioinformatics and Computational Biology, Las Vegas, NV, 2019.
- [90] A. Bax and N. Tjandra, "High-resolution heteronuclear NMR of human ubiquitin in an aqueous liquid crystalline medium.," *Journal of biomolecular NMR*, vol. 10, pp. 289-92, 1997.
- [91] A. Cupane, M. Leone, E. Vitrano, and L. Cordone, "Structural and dynamic properties of the heme pocket in myoglobin probed by optical spectroscopy.," *Biopolymers*, vol. 27, pp. 1977-97, 1988, doi: 10.1002/bip.360271209.
- [92] S. D. Emerson, J. T. J. Lecomte, and G. N. La Mar, "Proton NMR resonance assignment and dynamic analysis of phenylalanine CD1 in a low-spin ferric complex of sperm whale myoglobin," *Journal of the American Chemical Society*, vol. 110, pp. 4176-4182, 1988, doi: 10.1021/ja00221a013.
- [93] I. Bertini, C. Luchinat, P. Turano, G. Battaini, and L. Casella, "The magnetic properties of myoglobin as studied by NMR spectroscopy," *Chemistry-a European Journal*, vol. 9, pp. 2316-2322, 2003.
- [94] H. Shimada and W. S. Caughey, "Dynamic protein structures. Effects of pH on conformer stabilities at the ligand-binding site of bovine heart myoglobin carbonyl.," *The Journal of biological chemistry*, vol. 257, pp. 11893-900, 1982.
- [95] X. Huang *et al.*, "Replacement of Val3 in human thymidylate synthase affects its kinetic properties and intracellular stability ." *Biochemistry*, vol. 49, pp. 2475-2482, 2010, doi: 10.1021/bi901457e.
- [96] L. M. Gibson, L. L. Lovelace, and L. Lebioda, "The R163K mutant of human thymidylate synthase is stabilized in an active conformation: structural asymmetry and reactivity of cysteine 195.," *Biochemistry*, vol. 47, pp. 4636-4643, 2008, doi: 10.1021/bi7019386.
- [97] J. Feeney, B. Birdsall, N. V. Kovalevskaya, Y. D. Smurnyy, E. M. Navarro Peran, and V. I. Polshakov, "NMR structures of apo L. casei dihydrofolate reductase and its complexes with trimethoprim and NADPH: contributions to positive cooperative binding from ligand-induced refolding, conformational changes, and interligand hydrophobic interactions.," *Biochemistry*, vol. 50, pp. 3609-20, 2011.

- [98] M. R. Sawaya and J. Kraut, "Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: crystallographic evidence.," *Biochemistry*, vol. 36, pp. 586-603, 1997.
- [99] K. Umemoto, H. Leffler, A. Venot, H. Valafar, and J. H. Prestegard, "Conformational differences in liganded and unliganded states of Galectin-3.," *Biochemistry*, vol. 42, pp. 3688-95, 2003, doi: 10.1021/bi026671m.
- [100] M. Nowotny and W. Yang, "Structural and functional modules in RNA interference.," *Current opinion in structural biology*, vol. 19, pp. 286-93, 2009, doi: 10.1016/j.sbi.2009.04.006.
- [101] J. Cavanagh, W. J. Fairbrother, A. G. Palmer, M. Rance, and N. J. Skelton, "Protein NMR Spectroscopy: Principles and Practice, Second Edition," 2007.
- [102] G. Barbato, M. Ikura, L. E. Kay, R. W. Pastor, and A. Bax, "Backbone dynamics of calmodulin studied by nitrogen-15 relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible," *Biochemistry*, vol. 31, pp. 5269-5278, 1992, doi: 10.1021/bi00138a005.
- [103] J. L. Lorieu, J. M. Louis, and A. Bax, "Whole-body rocking motion of a fusion peptide in lipid bilayers from size-dispersed 15N NMR relaxation.," *Journal of the American Chemical Society*, vol. 133, pp. 14184-7, 2011, doi: 10.1021/ja2045309.
- [104] G. Lipari and A. Szabo, "Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity," *Journal of the American Chemical Society*, vol. 104, pp. 4546-4559, 1982, doi: 10.1021/ja00381a009.
- [105] G. Bouvignies *et al.*, "Identification of slow correlated motions in proteins using residual dipolar and hydrogen-bond scalar couplings," *Proc Natl Acad Sci U S A*, vol. 102, pp. 13885-13890, 2005.
- [106] G. M. Clore and C. D. Schwieters, "Amplitudes of protein backbone dynamics and correlated motions in a small alpha/beta protein: correspondence of dipolar coupling and heteronuclear relaxation measurements," *Biochemistry*, vol. 43, pp. 10678-10691, 2004.
- [107] J. R. Tolman, J. M. Flanagan, M. A. Kennedy, and J. H. Prestegard, "NMR evidence for slow collective motions in cyanometmyoglobin," *Nat Struct Biol*, vol. 4, pp. 292-297, 1997.
- [108] J. R. Tolman and K. Ruan, "NMR residual dipolar couplings as probes of biomolecular dynamics.," *Chemical reviews*, vol. 106, pp. 1720-1736, 2006, doi: 10.1021/cr040429z.



- [109] N. A. Lakomek, T. Carlomagno, S. Becker, C. Griesinger, and J. Meiler, "A thorough dynamic interpretation of residual dipolar couplings in ubiquitin," *J Biomol NMR*, vol. 34, pp. 101-115, 2006, doi: 10.1007/s10858-005-5686-0.
- [110] N.-A. Lakomek *et al.*, "Self-consistent residual dipolar coupling based model-free analysis for the robust determination of nanosecond to microsecond protein dynamics.," *Journal of biomolecular NMR*, vol. 41, pp. 139-55, 2008.
- [111] H. M. Al-Hashimi, Y. Gosser, A. Gorin, W. Hu, A. Majumdar, and D. J. Patel, "Concerted motions in HIV-1 TAR RNA may allow access to bound state conformations: RNA dynamics from NMR residual dipolar couplings.," *Journal of molecular biology*, vol. 315, pp. 95-102, 2002, doi: 10.1006/jmbi.2001.5235.
- [112] P. Bernadó and M. Blackledge, "Local dynamic amplitudes on the protein backbone from dipolar couplings: toward the elucidation of slower motions in biomolecules.," *Journal of the American Chemical Society*, vol. 126, pp. 7760-1, 2004, doi: 10.1021/ja048785m.
- [113] L. c. Salmon, G. Bouvignies, P. Markwick, and M. Blackledge, "Nuclear Magnetic Resonance Provides a Quantitative Description of Protein Conformational Flexibility on Physiologically Important Time Scales," *Biochemistry*, vol. 50, pp. 2735-2747, 2011, doi: 10.1021/bi200177v.
- [114] C. Luchinat, M. Nagulapalli, G. Parigi, and L. Sgheri, "Maximum occurrence analysis of protein conformations for different distributions of paramagnetic metal ions within flexible two-domain proteins," *Journal of Magnetic Resonance*, vol. 215, pp. 85-93, 2011, doi: 10.1016/j.jmr.2011.12.016.
- [115] A. D. Simone, R. W. Montalvao, and M. Vendruscolo, "Determination of Conformational Equilibria in Proteins Using Residual Dipolar Couplings," *Journal of Chemical Theory and Computation*, vol. 7, pp. 4189-4195, 2011.
- [116] M. Getz, X. Sun, A. Casiano-negrone, Q. Zhang, and H. M. Al-hashimi, "Rna dynamics and conformational adaptation," *Biopolymers*, vol. 86, pp. 384-402, 2007, doi: 10.1002/bip.
- [117] T. S. Ulmer, B. E. Ramirez, F. Delaglio, and A. Bax, "Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy," *J Am Chem Soc*, vol. 125, pp. 9179-9191, 2003, doi: 10.1021/ja0350684.
- [118] P. Guerry, L. Mollica, and M. Blackledge, "Mapping protein conformational energy landscapes using NMR and molecular simulation.," *Chemphyschem : a European journal of chemical physics and physical chemistry*, vol. 14, pp. 3046-58, 2013, doi: 10.1002/cphc.201300377.

- [119] A. De Simone, F. A. Aprile, A. Dhulesia, C. M. Dobson, and M. Vendruscolo, "Structure of a low-population intermediate state in the release of an enzyme product," *eLife*, vol. 4, p. e02777, 2015, doi: 10.7554/eLife.02777.
- [120] B. R. Brooks *et al.*, "CHARMM: the biomolecular simulation program.," *Journal of computational chemistry*, vol. 30, pp. 1545-1614, 2009, doi: 10.1002/jcc.21287.
- [121] R. Salomon-Ferrer, D. A. Case, and R. C. Walker, "An overview of the Amber biomolecular simulation package," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 3, pp. 198-210, 2013.
- [122] C. D. Schwieters, J. J. Kuszewski, N. Tjandra, and G. M. Clore, "The Xplor-NIH NMR molecular structure determination package.," *Journal of Magnetic Resonance*, vol. 160, pp. 65-73, 2003.
- [123] S. H. Park, W. S. Son, R. Mukhopadhyay, H. Valafar, and S. J. Opella, "Phage-induced alignment of membrane proteins enables the measurement and structural analysis of residual dipolar couplings with dipolar waves and lambda-maps.," *Journal of the American Chemical Society*, vol. 131, pp. 14140-1, 2009, doi: 10.1021/ja905640d.
- [124] L. Yao, B. Vögeli, D. A. Torchia, and A. Bax, "Simultaneous NMR study of protein structure and dynamics using conservative mutagenesis," *J Phys Chem B*, vol. 112, pp. 6045-6056, 2008.
- [125] A. T. Brünger *et al.*, "Crystallography & NMR system: A new software suite for macromolecular structure determination.," *Acta crystallographica. Section D, Biological crystallography*, vol. 54, pp. 905-21, 1998.
- [126] P. Güntert, "Automated NMR structure calculation with CYANA," *Methods Mol Biol*, vol. 278, pp. 353-378, 2004.
- [127] H. Valafar and J. H. Prestegard, "REDCAT: a residual dipolar coupling analysis tool.," *Journal of magnetic resonance (San Diego, Calif. : 1997)*, vol. 167, pp. 228-41, 2004, doi: 10.1016/j.jmr.2003.12.012.
- [128] G. Kummerlöwe and B. Luy, "Residual dipolar couplings as a tool in determining the structure of organic molecules," *TrAC Trends in Analytical Chemistry*, vol. 28, pp. 483-493, 2009.
- [129] C. M. Thiele, "Residual Dipolar Couplings (RDCs) in Organic Structure Determination," *European Journal of Organic Chemistry*, vol. 2008, pp. 5673-5685, 2008.
- [130] H. F. Azurmendi and C. A. Bush, "Conformational studies of blood group A and blood group B oligosaccharides using NMR residual dipolar couplings," *Carbohydrate Research*, vol. 337, pp. 905-915, 2002.

- [131] H. F. Azurmendi, M. Martin-Pastor, and C. A. Bush, "Conformational studies of Lewis X and Lewis A trisaccharides using NMR residual dipolar couplings," *Biopolymers*, vol. 63, pp. 89-98, 2002.
- [132] J. Adeyeye *et al.*, "Conformation of the hexasaccharide repeating subunit from the *Vibrio cholerae* O139 capsular polysaccharide," *Biochemistry*, vol. 42, pp. 3979-3988, 2003.
- [133] F. Tian, H. M. Al-Hashimi, J. L. Craighead, and J. H. Prestegard, "Conformational analysis of a flexible oligosaccharide using residual dipolar couplings," *Journal of the American Chemical Society*, vol. 123, pp. 485-492, 2001.
- [134] H. M. Al-Hashimi, A. Gorin, A. Majumdar, Y. Gossler, and D. J. Patel, "Towards structural Genomics of RNA: Rapid NMR resonance assignment and simultaneous RNA tertiary structure determination using residual dipolar couplings," *J Mol Biol*, vol. 318, pp. 637-649, 2002.
- [135] N. Tjandra, S. Tate, A. Ono, M. Kainosho, and A. Bax, "The NMR structure of a DNA dodecamer in an aqueous dilute liquid crystalline phase," *J. Am. Chem. Soc.*, vol. 122, pp. 6190-6200, 2000.
- [136] A. Vermeulen, H. Zhou, and A. Pardi, "Determining DNA Global Structure and DNA Bending by Application of NMR Residual Dipolar Couplings," *Journal of the American Chemical Society*, vol. 122, pp. 9638-9647, 2000, doi: 10.1021/ja001919l.
- [137] H. M. Al-Hashimi, P. J. Bolon, and J. H. Prestegard, "Molecular symmetry as an aid to geometry determination in ligand protein complexes," *J Magn Reson*, vol. 142, pp. 153-158, 2000.
- [138] G. Cornilescu, F. Delaglio, and A. Bax, "Protein backbone angle restraints from searching a database for chemical shift and sequence homology," *J Biomol NMR*, vol. 13, pp. 289-302, 1999.
- [139] C. A. Fowler, F. Tian, H. M. Al-Hashimi, and J. H. Prestegard, "Rapid determination of protein folds using residual dipolar couplings," *J Mol Biol*, vol. 304, pp. 447-460, 2000.
- [140] G. M. Clore and C. A. Bewley, "Using conjoined rigid body/torsion angle simulated annealing to determine the relative orientation of covalently linked protein domains from dipolar couplings," *J. Magn. Reson.*, vol. 154, pp. 329-335, 2002.
- [141] M. Assfalg, I. Bertini, P. Turano, A. Grant Mauk, J. R. Winkler, and H. B. Gray, "<sup>15</sup>N-<sup>1</sup>H Residual dipolar coupling analysis of native and alkaline-K79A *Saccharomyces cerevisiae* cytochrome c.," *Biophysical journal*, vol. 84, pp. 3917-23, 2003, doi: 10.1016/S0006-3495(03)75119-4.

- [142] J. H. Prestegard, H. M. Al-Hashimi, and J. R. Tolman, "NMR structures of biomolecules using field oriented media and residual dipolar couplings.," *Quarterly reviews of biophysics*, vol. 33, pp. 371-424, 2000.
- [143] R. J. Sanders, S. L. Hammond, and N. M. Rao, "Thoracic outlet syndrome: a review.," (in eng), *The neurologist*, vol. 14, pp. 365-73, 2008, doi: 10.1097/NRL.0b013e318176b98d.
- [144] J. H. Prestegard, C. M. Bougault, and A. I. Kishore, "Residual dipolar couplings in structure determination of biomolecules.," *Chemical reviews*, vol. 104, pp. 3519-40, 2004.
- [145] N. Tjandra, J. G. Omichinski, A. M. Gronenborn, G. M. Clore, and A. Bax, "Use of dipolar H-1-N-15 and H-1-C-13 couplings in the structure determination of magnetically oriented macromolecules in solution," *Nature Structural Biology*, vol. 4, pp. 732-738, 1997.
- [146] C. Guo, R. Godoy-Ruiz, and V. Tugarinov, "High resolution measurement of methyl  $^{13}\text{C}(m)$ - $^{13}\text{C}$  and  $^1\text{H}(m)$ - $^{13}\text{C}(m)$  residual dipolar couplings in large proteins.," *Journal of the American Chemical Society*, vol. 132, pp. 13984-7, 2010, doi: 10.1021/ja1041435.
- [147] J. a. Losonczi, M. Andrec, M. W. Fischer, and J. H. Prestegard, "Order matrix analysis of residual dipolar couplings using singular value decomposition.," *Journal of magnetic resonance (San Diego, Calif. : 1997)*, vol. 138, pp. 334-42, 1999, doi: 10.1006/jmre.1999.1754.
- [148] H. Valafar and J. H. Prestegard, "Rapid classification of a protein fold family using a statistical analysis of dipolar couplings.," *Bioinformatics (Oxford, England)*, vol. 19, pp. 1549-55, 2003.
- [149] X. Miao, R. Mukhopadhyay, and H. Valafar, "Estimation of relative order tensors, and reconstruction of vectors in space using unassigned RDC data and its application.," *Journal of magnetic resonance (San Diego, Calif. : 1997)*, vol. 194, pp. 202-11, 2008, doi: 10.1016/j.jmr.2008.07.005.
- [150] G. M. Clore, A. M. Gronenborn, and A. Bax, "A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information," *J Magn Reson*, vol. 133, pp. 216-221, 1998, doi: 10.1006/jmre.1998.1419.
- [151] M. Zweckstetter, "NMR: prediction of molecular alignment from structure using the PALES software," *Nat Protoc*, vol. 3, pp. 679-690, 2008.

- [152] J. J. Warren and P. B. Moore, "A maximum likelihood method for determining D(a)(PQ) and R for sets of dipolar coupling data.," *Journal of magnetic resonance (San Diego, Calif. : 1997)*, vol. 149, pp. 271-5, 2001, doi: 10.1006/jmre.2001.2307.
- [153] L. L. Lovelace, S. R. Johnson, L. M. Gibson, B. J. Bell, S. H. Berger, and L. Lebioda, "Variants of human thymidylate synthase with loop 181-197 stabilized in the inactive conformation.," *Protein science : a publication of the Protein Society*, vol. 18, pp. 1628-36, 2009, doi: 10.1002/pro.171.
- [154] N. A. Greshenfeld, "The Nature of Mathematical Modeling," 1998.
- [155] K. Levenberg, "A method for the solution of certain problems in least squares," *Quarterly of Applied Mathematics*, vol. 2, pp. 164 - 168, 1944.
- [156] S. Y. Sohn, W. J. Bae, J. J. Kim, K.-H. Yeom, V. N. Kim, and Y. Cho, "Crystal structure of human DGCR8 core.," *Nature structural & molecular biology*, vol. 14, pp. 847-53, 2007.
- [157] C. Wostenberg, W. G. Noid, and S. A. Showalter, "MD simulations of the dsRBP DGCR8 reveal correlated motions that may aid pri-miRNA binding.," (in English), *Biophysical journal*, vol. 99, pp. 248-56, 2010.
- [158] S. Wu, J. Skolnick, and Y. Zhang, "Ab initio modeling of small proteins by iterative TASSER simulations," *BMC Biol*, vol. 5, p. 17, 2007.
- [159] Y. Zhang, "Template-based modeling and free modeling by I-TASSER in CASP7," *Proteins*, vol. 69 Suppl 8, pp. 108-117, 2007.
- [160] H. M. Al-Hashimi, H. Valafar, M. Terrell, E. R. Zartler, M. K. Eidsness, and J. H. Prestegard, "Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings.," *Journal of magnetic resonance (San Diego, Calif. : 1997)*, vol. 143, pp. 402-6, 2000, doi: 10.1006/jmre.2000.2049.
- [161] R. Koradi, M. Billeter, K. Wuthrich, and K. Wuthrich, "MOLMOL: A program for display and analysis of macromolecular structures," *J Mol Graphics*, vol. 14, pp. 51-55, 1996, doi: 10.1016/0263-7855(96)00009-4.
- [162] J. R. Tolman, "A novel approach to the retrieval of structural and dynamic information from residual dipolar couplings using several oriented media in biomolecular NMR spectroscopy.," *Journal of the American Chemical Society*, vol. 124, pp. 12020-30, 2002.
- [163] W. H. Organization, "Warning about the dangers of tobacco," *WHO Report on the global tobacco epidemic*, vol. 152, 2011.

- [164] CDC, "Annual Smoking-Attributable Mortality, Years of Potential Life Lost, and Productivity Losses—United States, 1995–1999," *MMWR*, vol. 51, no. 14, pp. 300-313, November 13, 2009 2002.
- [165] B. Bonevski, J. Bryant, M. Lynagh, and C. Paul, "Money as motivation to quit: a survey of a non-random Australian sample of socially disadvantaged smokers' views of the acceptability of cash incentives," *Prev Med*, vol. 55, no. 2, pp. 122-6, Aug 2012, doi: 10.1016/j.ypmed.2012.06.001.
- [166] D. L. Patrick, A. Cheadle, D. C. Thompson, P. Diehr, T. Koepsell, and S. Kinne, "The validity of self-reported smoking: a review and meta-analysis.," *American Journal of Public Health*, vol. 84, pp. 1086-1093, 1994, doi: 10.2105/AJPH.84.7.1086.
- [167] P. R. Center, "Share of adults in the United States who owned a smartphone from 2011 to 2017, by location," 2018.
- [168] C. T. Association, "Smartwatch devices unit sales in the United States from 2013 to 2017 (in millions)," 2016.
- [169] C. T. Association, "Smartwatch unit sales worldwide from 2014 to 2018 (in millions)," 2017.
- [170] V. Lambert-Jessup, S. Ferguson, F. Islam, D. Hammond, J. Niederdeppe, and J. Thrasher, "Assessing the impact of supportive health messages on cigarette package inserts: A pilot study using ecological momentary assessment.," in *Society for Research on Nicotine & Tobacco*, Baltimore, MD, 2018.
- [171] L. E. Wagenknecht, G. L. Burke, L. L. Perkins, N. J. Haley, and G. D. Friedman, "Misclassification of smoking status in the CARDIA study: a comparison of self-report with serum cotinine levels.," *American Journal of Public Health*, vol. 82, pp. 33-36, 1992, doi: 10.2105/AJPH.82.1.33.
- [172] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment.," *Annual review of clinical psychology*, vol. 4, pp. 1-32, 2008, doi: 10.1146/annurev.clinpsy.3.022806.091415.
- [173] A. Parate, M.-C. Chiu, C. Chadowitz, D. Ganesan, and E. Kalogerakis, "RisQ: Recognizing Smoking Gestures with Inertial Sensors on a Wristband.," *MobiSys ... : the ... International Conference on Mobile Systems, Applications and Services. International Conference on Mobile Systems, Applications, and Services*, vol. 2014, pp. 149-161, 2014.
- [174] N. Saleheen *et al.*, New York, New York, USA. puffMarker.

- [175] A. L. Skinner, C. J. Stone, H. Doughty, and M. R. Munafo, "StopWatch: The Preliminary Evaluation of a Smartwatch-Based System for Passive Detection of Cigarette Smoking," *Nicotine Tob Res*, vol. 21, no. 2, pp. 257-261, Jan 4 2019, doi: 10.1093/ntr/nty008.
- [176] R. Dar, "Effect of Real-Time Monitoring and Notification of Smoking Episodes on Smoking Reduction: A Pilot Study of a Novel Smoking Cessation App," *Nicotine Tob Res*, vol. 20, no. 12, pp. 1515-1518, Nov 15 2018, doi: 10.1093/ntr/ntx223.
- [177] K. Li *et al.*, "Smoking and Risk of All-cause Deaths in Younger and Older Adults: A Population-based Prospective Cohort Study Among Beijing Adults in China.," *Medicine*, vol. 95, p. e2438, 2016.
- [178] B. L. Rooney, P. Silha, J. Gloyd, and R. Kreutz, "Quit and Win smoking cessation contest for Wisconsin college students.," *WMJ : official publication of the State Medical Society of Wisconsin*, vol. 104, pp. 45-9, 2005.
- [179] I. Khati, G. Menvielle, A. Chollet, N. Younès, B. Metadieu, and M. Melchior, "What distinguishes successful from unsuccessful tobacco smoking cessation? Data from a study of young adults (TEMPO)." *Preventive medicine reports*, vol. 2, pp. 679-85, 2015.
- [180] D. W. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, pp. 431-441, 1963, doi: 10.1017/CBO9781107415324.004.
- [181] A. Ranganathan, "The Levenberg-Marquardt Algorithm," ed, 2004.
- [182] N. Schuz, M. Eid, B. Schuz, and S. G. Ferguson, "Immediate effects of plain packaging health warnings on quitting intention and potential mediators: Results from two ecological momentary assessment studies.," *Psychology of Addictive Behaviors*, vol. 30, pp. 220-228, 2016, doi: 10.1037/adb0000146.
- [183] E. T. Hebert, E. A. Vandewater, M. S. Businelle, M. B. Harrell, S. H. Kelder, and C. L. Perry, "Feasibility and reliability of a mobile tool to evaluate exposure to tobacco product marketing and messages using ecological momentary assessment," *Addictive Behaviors*, vol. 73, pp. 105-110, 2017, doi: 10.1016/j.addbeh.2017.05.004.
- [184] L. E. Burke *et al.*, "Ecological Momentary Assessment in Behavioral Research: Addressing Technological and Human Participant Challenges.," *Journal of medical Internet research*, vol. 19, p. e77, 2017, doi: 10.2196/jmir.7138.
- [185] A. Stone, S. Shiffman, A. Atienza, and L. Nebeling, "The Science of Real-Time Data Capture: Self-Reports in Health Research," p. 416, 2007.

- [186] L.-C. L.-H. Wu, L.-C. L.-H. Wu, and S.-C. Chang, "Exploring consumers' intention to accept smartwatch," *Computers in Human Behavior*, vol. 64, pp. 383-392, 2016, doi: 10.1016/j.chb.2016.07.005.
- [187] T. Prioleau, E. Moore, and M. Ghovanloo, "Unobtrusive and Wearable Systems for Automatic Dietary Monitoring," *IEEE Transactions on Biomedical Engineering*, pp. 1-1, 2017, doi: 10.1109/TBME.2016.2631246.
- [188] G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber, Las Vegas, NV. Smartwatch-based Activity Recognition: A Machine Learning Approach.
- [189] O. Akyazi, S. Batmaz, B. Kosucu, and B. Arnrich. SmokeWatch: A smartwatch smoking cessation assistant.
- [190] B. Bhandari, J. Lu, X. Zheng, S. Rajasegarar, and C. Karmakar. Non-invasive sensor based automated smoking activity detection.
- [191] E. Sazonov, P. Lopez-Meyer, and S. Tiffany, "A wearable sensor system for monitoring cigarette smoking.," *Journal of studies on alcohol and drugs*, vol. 74, pp. 956-64, 2013.
- [192] C. A. Cole, B. Janos, D. Anshari, J. F. Thrasher, S. Strayer, and H. Valafar, "Recognition of Smoking Gesture Using Smart Watch Technology," in *Proceedings of the International Conference on Health Informatics and Medical Systems (HIMS)*, Las Vegas, NV USA, 2016.
- [193] C. A. Cole, J. F. Thrasher, S. Strayer, and H. Valafar, "Resolving Ambiguities in Accelerometer Data Due to Location of Sensor on Wrist in Application to Detection of Smoking Gesture," in *IEEE International Conference on Biomedical and Health Informatics*, Orlando, FL, USA 16-19 Feb. 2017 2017: IEEE, pp. 489-492, doi: 10.1109/BHI.2017.7897312
- [194] E. T. Middleton and A. H. Morice, "Breath carbon monoxide as an indication of smoking habit.," *Chest*, vol. 117, pp. 758-63, 2000.
- [195] M. S. Pearce and L. Hayes, "Self-Reported Smoking Status and Exhaled Carbon Monoxide," *Chest*, vol. 128, pp. 1233-1238, 2005, doi: 10.1378/chest.128.3.1233.
- [196] A. Sandberg, C. M. Sköld, J. Grunewald, A. Eklund, and Å. M. Wheelock, "Assessing Recent Smoking Status by Measuring Exhaled Carbon Monoxide Levels," *PLoS ONE*, vol. 6, p. e28864, 2011, doi: 10.1371/journal.pone.0028864.
- [197] D. E. Rumelhart and J. L. McClelland, "Parallel distributed processing: explorations in the microstructure of Cognition. Volume 1. Foundations," *Nature*, vol. 327, p. 564, 1986.



- [198] AHA *et al.*, "Heart Disease and Stroke Statistics--2010 Update: A Report From the American Heart Association," *Circulation*, vol. 121, no. 7, pp. e46-215, February 23, 2010, doi: 10.1161/circulationaha.109.192667.
- [199] M. Shoaib, S. Bosch, H. Scholten, P. J. M. Havinga, and O. D. Incel. Towards detection of bad habits by fusing smartphone and smartwatch sensors.
- [200] M. Shoaib, S. Bosch, D. O. Incel, H. Scholten, and J. P. Havinga, "Complex Human Activity Recognition Using Smartphone and Wrist-Worn Motion Sensors," *Sensors*, vol. 16, no. 4, 2016, doi: 10.3390/s16040426.
- [201] C. A. Cole, D. Anshari, V. Lambert, J. F. Thrasher, and H. Valafar, "Detecting Smoking Events Using Accelerometer Data Collected Via Smartwatch Technology: Validation Study," *JMIR mHealth and uHealth*, vol. 5, p. e189, 2017, doi: 10.2196/mhealth.9035.
- [202] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher. Activity Recognition and Monitoring Using Multiple Sensors on Different Body Positions.
- [203] C. A. C. Chrisogonas Odhiambo, Alaleh Torkjazi, Homayoun Valafar, "State Transition Modeling of the Smoking Behavior using LSTM Recurrent Neural Networks," in *Computational Science & Computational Intelligence USA*, Las Vegas, L. D. H. R. Arabnia, F. G. Tinetti, Ed., 2019, IEEE Xplore: IEEE Xplore.
- [204] C. A. C. Homayoun Valafar, James Thrasher, Scott Strayer, "Detecting Smoking Gestures Using Accelerometer Data and Artificial Neural Networks," 2017.
- [205] M. Shoaib, H. Scholten, P. J. M. Havinga, and O. D. Incel. A hierarchical lazy smoking detection algorithm using smartwatch sensors.
- [206] E. M. Lee, J. L. Malson, A. J. Waters, E. T. Moolchan, and W. B. Pickworth, "Smoking topography: reliability and validity in dependent smokers," *Nicotine Tob Res*, vol. 5, no. 5, pp. 673-9, Oct 2003, doi: 10.1080/1462220031000158645.
- [207] M. D. Blank, S. Disharoon, and T. Eissenberg, "Comparison of methods for measurement of smoking behavior: mouthpiece-based computerized devices versus direct observation," *Nicotine Tob Res*, vol. 11, no. 7, pp. 896-903, Jul 2009, doi: 10.1093/ntr/ntp083.
- [208] K. A. Perkins and J. L. Karelitz, "A Procedure to Standardize Puff Topography During Evaluations of Acute Tobacco or Electronic Cigarette Exposure," *Nicotine Tob Res*, Dec 8 2018, doi: 10.1093/ntr/nty261.
- [209] B. Froeliger *et al.*, "Association Between Baseline Corticothalamic-Mediated Inhibitory Control and Smoking Relapse Vulnerability," *JAMA Psychiatry*, vol. 74, no. 4, pp. 379-386, Apr 1 2017, doi: 10.1001/jamapsychiatry.2017.0017.

- [210] C. A. Conklin *et al.*, "Combined Smoking Cues Enhance Reactivity and Predict Immediate Subsequent Smoking," (in eng), *Nicotine Tob Res*, vol. 21, no. 2, pp. 241-248, Jan 4 2019, doi: 10.1093/ntr/nty009.
- [211] E. Childs and H. de Wit, "Effects of acute psychosocial stress on cigarette craving and smoking," (in eng), *Nicotine Tob Res*, vol. 12, no. 4, pp. 449-53, Apr 2010, doi: 10.1093/ntr/ntp214.
- [212] A. H. Weinberger and S. A. McKee, "Gender differences in smoking following an implicit mood induction," *Nicotine & Tobacco Research*, vol. 14, no. 5, pp. 621-625, 2012.
- [213] R. F. Leeman, S. S. O'Malley, M. A. White, and S. A. McKee, "Nicotine and food deprivation decrease the ability to resist smoking," (in eng), *Psychopharmacology*, vol. 212, no. 1, pp. 25-32, 2010, doi: 10.1007/s00213-010-1902-z.
- [214] E. A. McClure *et al.*, "The influence of gender and oxytocin on stress reactivity, cigarette craving, and smoking in a randomized, placebo-controlled laboratory relapse paradigm," (in eng), *Psychopharmacology*, Dec 3 2019, doi: 10.1007/s00213-019-05392-z.
- [215] S. A. McKee, A. H. Weinberger, J. Shi, J. Tetrault, and S. Coppola, "Developing and validating a human laboratory model to screen medications for smoking cessation," (in eng), *Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco*, vol. 14, no. 11, pp. 1362-1371, 2012, doi: 10.1093/ntr/nts090.
- [216] R. L. Tomko *et al.*, "Methods to reduce the incidence of false negative trial results in substance use treatment research," (in eng), *Curr Opin Psychol*, vol. 30, pp. 35-41, Dec 2019, doi: 10.1016/j.copsyc.2019.01.009.
- [217] N. L. Benowitz *et al.*, "Biochemical Verification of Tobacco Use and Abstinence: 2019 Update," (in eng), *Nicotine Tob Res*, Oct 1 2019, doi: 10.1093/ntr/ntz132.
- [218] M. E. Piper *et al.*, "Defining and measuring abstinence in clinical trials of smoking cessation interventions: An updated review," (in eng), *Nicotine Tob Res*, Jul 4 2019, doi: 10.1093/ntr/ntz110.
- [219] M. S. Businelle, P. Ma, D. E. Kendzor, S. G. Frank, D. J. Vidrine, and D. W. Wetter, "An Ecological Momentary Intervention for Smoking Cessation: Evaluation of Feasibility and Effectiveness," (in eng), *J Med Internet Res*, vol. 18, no. 12, p. e321, Dec 12 2016, doi: 10.2196/jmir.6058.
- [220] Y. Cho, Y. Nam, Y.-J. Choi, and W.-D. Cho, "SmartBuckle," *Proceedings of the 2nd International Workshop on Systems and Networking Support for Health Care*

- and Assisted Living Environments - HealthNet '08*, p. 1, 2008, doi: 10.1145/1515747.1515757.
- [221] J. Yang, "Toward Physical Activity Diary : Motion Recognition Using Simple Acceleration Features with Mobile Phones," *Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics*, pp. 1-9, 2009, doi: 10.1145/1631040.1631042.
- [222] P. M. Scholl and K. van Laerhoven. A Feasibility Study of Wrist-Worn Accelerometer Based Detection of Smoking Habits.

## Appendix A: Discussion of Decimation

### Pseudocode of previous version of REDCRAFT's decimation routine

Given: *Search\_Depth*: <integer> Number of structures to be propagated forward  
*Score\_Threshold*: <float> Threshold to filter remaining angles (angstroms)  
*Cluster\_Sensitivity*: <float> Defines the sensitivity of clusters (angstroms)  
*Possible\_Angles*: <float> m by 2n matrix where n is current residue number and m represents the total search space for the addition of a new residue.

Function Decimation(*Possible\_Angles*)

*New\_Angles* = *Possible\_Angles*[1:*Search\_Depth*, :]  
*Remaining\_Angles* = *Possible\_Angles*[*Search\_Depth* + 1:end, :]  
*Clusters* = Cluster(*Remaining\_Angles*, *Cluster\_Sensitivity*)  
For each *Cluster c* in *Clusters* s.t. *c.score* <= *Score\_Threshold*  
    *Centroid* = ExtractCenterOfCluster(*c*)  
    Add *Centroid* to *New\_Angles*  
Return *New\_Angles*

### Updates made to the Pseudocode (in bold)

Given: *Search\_Depth*, *Cluster\_Sensitivity*, *Possible\_Angles*: same as before  
*Score\_Threshold*: <float> **percentage increase from bottom score of New\_Angles array (ranging from 0-1)**  
*Max\_Structures*: <int> **limit on the maximum structures extracted**

Function Decimation(*Possible\_Angles*)

*New\_Angles* = *Possible\_Angles*[1:*Search\_Depth*, :]  
*Max\_Score* = *New\_Angles*[end,:].score  
*Remaining\_Angles* = *Possible\_Angles*[*Search\_Depth* + 1:end, :]  
*Clusters* = Cluster(*Remaining\_Angles*, *Cluster\_Sensitivity*)  
**Sort Clusters based on score (smallest to largest)**  
**For each Cluster c in Clusters in the range 1:Max\_Structures**  
    Break if *c.score* > ***Max\_Score* \* *Score\_Threshold***  
    *Centroid* = ExtractCenterOfCluster(*c*)  
    Add *Centroid* to *New\_Angles*  
Return *New\_Angles*

## Appendix B: List of Publications

### Patents

1. Homayoun Valafar, Casey A. Cole, James F. Thrasher, Scott M. Strayer. Wearable computing device featuring machine-learning-based smoking detection. 2020 Patent #: US10551935.

### Journal Publications

2. Casey A. Cole, Nourhan S. Daigham, Gaohua Liu, Gaetano T. Montelione, Homayoun Valafar. Structure Determination of Proteins Using Residual Dipolar Coupling Data from Perdeuterated Proteins. In preparation 2020.
3. Casey A. Cole, Shannon Powers, Rachel Tomko, Brett Froeliger and Homayoun Valafar. (2020) Clinical Quantification of Smoking Topography Using Smartwatch Technology. Submitted, awaiting decision.
4. Casey A. Cole, Caleb Parks, Julian Rachele and Homayoun Valafar, Increased Usability, Algorithmic Improvements and Incorporation of Data Mining for Structure Calculation of Proteins with REDCRAFT Software Package, (2020) Accepted BMC Bioinformatics, awaiting publication.
5. Annalisa M Baratta, Nickole R Kanyuch, Casey A Cole, Homayoun Valafar, Jessica Deslauriers, Ana Pocivavsek. Acute Sleep Deprivation During Pregnancy in Rats: Rapid Elevation of Placental and Fetal Inflammation and Kynurenic Acid. *Neurobiology of Stress*, 2020 CiteScore: 8.28, SNIP: 2.094, SJR: 3.359.
6. Casey A. Cole, Dien Anshari, Victoria Lambert, James F. Thrasher, Homayoun Valafar (2017) Detecting Smoking Events Using Accelerometer Data Collected Via Smartwatch Technology: A Feasibility Study. *JMIR mHealth and uHealth*, 5(12):e189. DOI: 10.2196/mhealth.9035
7. Casey A. Cole, Rishi Mukhopadhyay, Hanin Omar, Mirko Hennig, Homayoun Valafar. (2016) Structure Calculation and Reconstruction of Discrete State Dynamics from Residual Dipolar Couplings. *Journal of Chemical Theory and Computation*, 12 (4), pp 1408–1422. DOI: 10.1021/acs.jctc.5b01091
8. Simin, Mikhail, Irausquin, Stephanie, Cole, Casey A., & Valafar, Homayoun. Improvements to REDCRAFT: a software tool for simultaneous characterization

of protein backbone structure and dynamics from Residual Dipolar Couplings. *Journal of Biomolecular NMR*, Volume 60, Issue 4, pp. 241-264, December **2014**.

#### Book Chapters

9. C.A. Cole, D. Ishimaru, M. Hennig, H. Valafar. (**2016**). Structure Calculation of  $\alpha$ ,A/ $\beta$ ,  $\beta$  Proteins From Residual Dipolar Coupling Data Using Redcraft. In: H. R. Arabnia, Q. N. Tran (Eds.), *Emerging Trends in Applications and Infrastructures for Computational Biology, Bioinformatics, and Systems Biology*. Morgan Kaufmann, Imprint of Elsevier, Cambridge, MA, pp. 73–88

#### Conference Publications

10. Casey A. Cole, Christopher Ott, Diego Valdes, Homayoun Valafar; "PDBMine: A Reformulation of the Protein Data Bank to Facilitate Structural Data Mining"; Proceedings of 2019 International Conference on Computational Science & Computational Intelligence (Chapter: Symposium on Computational Biology); CSCI, USA, Co-Editors: H. R. Arabnia, L. Deligiannidis, and F. G. Tinetti; IEEE CPS, IEEE Xplore entry: 1803739, December 2019.
11. Chrisogonas Odhiambo\*, Casey A. Cole, Alaleh Torkjazi, and Homayoun Valafar; State Transition Modeling of the Smoking Behavior using LSTM Recurrent Neural Networks; Proceedings of 2019 International Conference on Computational Science & Computational Intelligence (Chapter: Health Informatics & Medical Systems); CSCI, USA, Co-Editors: H. R. Arabnia, L. Deligiannidis, and F. G. Tinetti; IEEE CPS. December 2019
12. Casey A. Cole, Caleb Parks, Julian Rachele and Homayoun Valafar, *Improvements of the REDCRAFT Software Package*, Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP), July 2019, Las Vegas, NV. Acceptance Rate: 29%
13. Casey A. Cole, Kenneth L. Nesbitt, Homayoun Valafar, *Application of Ensemble Learning to the Differential Gene Expression in Left-Right Breast Tumor*, Contributed paper at 2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), May 2018, St. Louis, MO.
14. Casey A. Cole, James F. Thrasher, Scott Strayer, Homayoun Valafar, *Resolving Ambiguities in Accelerometer Data Due to Location of Sensor on Wrist in Application to Detection of Smoking Gesture*, Contributed paper at 2017 IEEE International Conference on Biomedical and Health Informatics, February **2017**, Orlando, Fl. 38% acceptance rate.
15. Casey A. Cole, Bethany Janos, Dien Anshari, James F. Thrasher, Scott Strayer, Homayoun Valafar, *Recognition of Smoking Gesture Using Smart Watch*

- Technology*, Proceedings of the International Conference on Health Informatics and Medical Systems (HIMS), July **2016**, Las Vegas, NV.
16. Casey A. Cole, Mirko Hennig, Homayoun Valafar. *An Investigation of Minimum Data Requirement for Successful Structure Determination of Pf2048.1 with REDCRAFT*, Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP), July **2015**, Las Vegas, NV.
  17. Hanin Omar, Casey A. Cole, Arjang Fahim, Giuliana Gusmaroli, Stephen Borgianini, Homayoun Valafar. *De Novo Assembly of Uca minax Transcriptome from Next Generation Sequencing*, Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP), July **2015**, Las Vegas, NV.

#### Oral Presentations

18. Casey A. Cole, Bethany Janos, Dien Anshari, James F. Thrasher, Scott Strayer, Homayoun Valafar, *Recognition of Smoking Gesture Using Smart Watch Technology*, Oral Presentation at the International Conference on Health Informatics and Medical Systems (HIMS), July **2016**, Las Vegas, NV.
19. Casey A. Cole, Bethany Janos, Dien Anshari, Mark Macaуда, Jim Thrasher, Scott Strayer, Homayoun Valafar, *Detecting Smoking Gestures Using Accelerometer Data and Artificial Neural Networks*, Oral presentation at National IDeA Symposium of Biomedical Research Excellence (NISBRE), June **2016**, Washington D.C.
20. Kenneth Nesbitt, Casey A. Cole, Nick Boltin, Yuxin Cui, Hao Ji, Min Li, Zhenhua Shang, Hanin Omar, Arjang Fahim, Misha Shtutman, Homayoun Valafar, *From Zero to Gene Assembly in Three Weeks*, Oral presentation at National IDeA Symposium of Biomedical Research Excellence (NISBRE), June 2016, Washington D.C.
21. Casey A. Cole, Mirko Hennig, Homayoun Valafar. *An Investigation of Minimum Data Requirement for Successful Structure Determination of Pf2048.1 with REDCRAFT*, 15 minute oral presentation at the International Conference on Bioinformatics & Computational Biology (BIOCOMP), July **2015**, Las Vegas, NV.

#### Poster Presentations

22. Casey A. Cole, Meng Zhang, Xiaowei Yu, Yan Xu, Gaetano T. Montelione, and Homayoun Valafar, *Reconstruction of Discrete State Dynamics Using Residual Dipolar Couplings*, 2017 SC INBRE Symposium, October **2017**, Columbia, SC.

23. Casey A. Cole, Kenneth Nesbitt, Homayoun Valafar, Using Genetic Sequences from Freely Available Databases as Input to Multiple, Orthogonal Algorithms for Gene Expression shows Genetic Profile Differences Between Left and Right Human Breast Tumors, 2017 SC INBRE Symposium, October **2017**, Columbia, SC.
24. Caleb Parks, Casey A. Cole, Homayoun Valafar, Structure Calculation of Proteins From NMR Data, 2017 SC INBRE Symposium, October **2017**, Columbia, SC.
25. Casey A. Cole, Meng Zhang, Xiaowei Yu, Yan Xu, Gaetano T. Montelione, and Homayoun Valafar, Reconstruction of Discrete State Dynamics Using Residual Dipolar Couplings , Gordon Research Conference on Computational Aspects of NMR, June **2017**, Newry, ME.
26. Yan, Chunmei, Gomez, Belinda, Hernandez Gifford, Jennifer, Cole, Casey, and LaVoie, Holly A. *C/EBP $\beta$  targeting to putative amino acid response elements in genes regulating ovarian function*. Carolina Women's Health Research Forum. University of South Carolina. Columbia, SC, Abstract 52, Nov. 4, **2016**.
27. Casey A. Cole, Hanin Omar, Homayoun Valafar. *An Exploration of the Energy Landscape of Dynamical Proteins Using NMR Data*, Gamecock Computing Research Symposium (poster presentation), November **2016**, Columbia, SC
28. Casey A. Cole, Bethany Janos, Dien Anshari, Mark Macaуда, Jim Thrasher, Scott Strayer, Homayoun Valafar. *Detecting Smoking Gestures Using Accelerometer Data*, SC INBRE Symposium (poster presentation), August **2016**, Columbia, SC
29. Arjang Fahim, Casey A. Cole, Homayoun Valafar. *nDPDPA: A Novel Hybrid Refinement Method Using Residual Dipolar Coupling*, SC INBRE Symposium (poster presentation), August **2016**, Columbia, SC.
30. Kenneth Nesbitt, Casey A. Cole, Nick Boltin, Yuxin Cui, Hao Ji, Min Li, Zhenhua Shang, Hanin Omar, Arjang Fahim, Misha Shtutman, Homayoun Valafar, *From Zero to Gene Assembly in Three Weeks* , SC INBRE Symposium (poster presentation), August **2016**, Columbia, SC.
31. Hanin Omar, Casey A. Cole, Mirko Hennig, Homayoun Valafar. *Characterization of Discrete State Dynamics from Residual Dipolar Couplings using REDCRAFT*, SC INBRE Symposium (poster presentation), August **2016**, Columbia, SC.
32. Arjang Fahim, Casey A. Cole, Homayoun Valafar. *nDPDPA: A Novel Hybrid Refinement Method Using Residual Dipolar Coupling*, ENC (accepted poster presentation), April **2016**, Pittsburgh, PA



33. Hanin Omar, Casey A. Cole, Mirko Hennig, Homayoun Valafar. *Characterization of Discrete State Dynamics from Residual Dipolar Couplings using REDCRAFT*, ENC (accepted poster presentation), April **2016**, Pittsburgh, PA
34. Casey A. Cole, Mirko Hennig, Homayoun Valafar. *Structure Elucidation of a Novel Protein from Minimal Set of Backbone RDC Data with REDCRAFT*, SC INBRE Symposium (poster presentation), February **2015**, Columbia, SC
35. Casey A. Cole, Hanin Omar, Arjang Fahim, Guiliana Gusmaroli, Homayoun Valafar. *Transcriptome Assembly of the Uca minax*, Poster presented at National IDeA Symposium of Biomedical Research Excellence (NISBRE), June **2014**, Washington D.C.
36. Cole, Casey A., Omar, Hanin, Fahim, Arjang, Gusmaroli, Giuliana & Valafar, Homayoun (November **2013**). *Database of Uca minax Transcriptome*. Poster session presented at the SE Regional IdeA Meeting, Little Rock, Arkansas.
37. Cole, Casey A., Buell, Duncan, Cooley, Heidi (April **2014**). *3D Modeling and Art Assets for Ghosts of the Horseshoe*. Poster session presented at the USC Discovery Day, Columbia, SC.
38. Cole, Casey A., Omar, Hanin, Fahim, Arjang, Gusmaroli, Giuliana & Valafar, Homayoun (June **2014**). *Transcriptome Assembly of Uca minax*. Poster session presented at the National IDeA Symposium of Biomedical Research Excellence (NISBRE), Washington, DC.
39. Omar, Hanin, Cole, Casey A., Fahim, Arjang, Gusmaroli, Giuliana & Valafar, Homayoun (October **2014**). *Transcriptome Assembly of Uca minax*. Poster session presented at the Gamecock Computing Symposium 2014, Columbia, SC.

## Appendix C: Permission for Reprint

*Publication reported in Section 2.1:* Casey A. Cole, Caleb Parks, Julian Rachele and Homayoun Valafar. Accepted by BMC Bioinformatics. Reprinted here with permission of publisher under Creative Commons Attribution License 4.0 as shown on the BMC biomedcentral website:

biomedcentral.com/about/policies/license-agreement

In the following, we provide the licenses' summaries as they can be found on the Creative Commons website.

The [Creative Commons Attribution License 4.0](#) provides the following summary (where 'you' equals 'the user'):

### You are free to:

- Share — copy and redistribute the material in any medium or format.
- Adapt — remix, transform, and build upon the material for any purpose, even commercially.

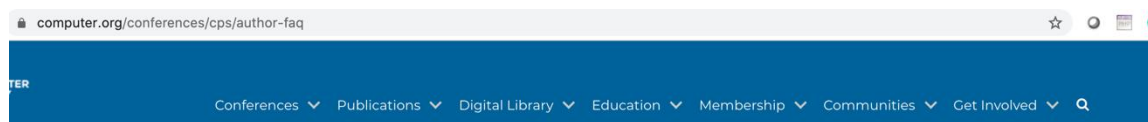
The licensor cannot revoke these freedoms as long as you follow the license terms.

### Under the following terms:

- Attribution— you must give *appropriate credit*, provide a link to the license, and *indicate if changes were made*. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- No additional restrictions—you may not apply legal terms or *technological measures* that legally restrict others from doing anything the license permits.

*Publication reported in Section 2.2:* Casey A. Cole, Christopher Ott, Diego Valdes and Homayoun Valafar. 2019. Proceedings of 2019 IEEE International Conference on Computational Science & Computational Intelligence (IEEE CSCI). Reprinted here with permission from the publisher as shown on the following website maintained by IEEE. It states that the work can be reposted anywhere without permission as long as the reader is

directed to the IEEE copyright policy page:  
<https://www.ieee.org/publications/rights/copyright-policy.html>.



^ How do I submit my final paper PDF file?

v Who owns the paper's copyright?

If the conference content is under IEEE copyright, it is required that authors or their employers transfer copyright to IEEE, except for papers in the public domain or those that are reprinted with permission from a previously published, copyrighted publication. After transferring copyright to IEEE, authors and/or their companies may post their IEEE-copyrighted material on their websites without permission, provided that the sites alert readers to their obligations with respect to copyrighted material and that the posted work includes an IEEE copyright notice. [Please refer to the IEEE Copyright Policy page for more information.](#)

- **Does IEEE require individuals working on a thesis or dissertation to obtain formal permission for reuse?**

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you must follow the requirements listed below:

#### **Textual Material**

Using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.

In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.

If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

#### **Full-Text Article**

If you are using the entire IEEE copyright owned article, the following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]

Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.

*Publication reported in Section 2.3:* Casey A. Cole, Rishi Mukhopadhyay, Hanin Omar, Mirko Hennig, and Homayoun Valafar. 2016. J. Chem. Theory Comput., 12 (4), pp 1408–1422. The following has been reprinted with permission from the publisher (see image below):



## Use in Theses/Dissertations

The following wording is from the ACS Thesis/Dissertation Policy and the ACS Journal Publishing Agreement:

**Reuse/Republication of the Entire Work in Theses or Collections:** Authors may reuse all or part of the Submitted, Accepted or Published Work in a thesis or dissertation that the author writes and is required to submit to satisfy the criteria of degree-granting institutions. Such reuse is permitted subject to the ACS' "[Ethical Guidelines to Publication of Chemical Research](#)"; the author should secure written confirmation (via letter or email) from the respective ACS journal editor(s) to avoid potential conflicts with journal prior publication\*/embargo policies. Appropriate citation of the Published Work must be made\*\*. If the thesis or dissertation to be published is in electronic format, a direct link to the Published Work must also be included using the [ACS Articles on Request](#) author-directed link.

\* Prior publication policies of ACS journals are posted on the [ACS website](#).

\*\* "Reprinted with permission from [COMPLETE REFERENCE CITATION]. Copyright [YEAR] American Chemical Society." Insert the appropriate wording in place of the capitalized words. This credit line wording should appear **on the first page of your ACS journal article**.

If your university requires written permission and your manuscript has not yet received a DOI (published ASAP), send a request to [copyright@acs.org](mailto:copyright@acs.org) that includes the manuscript number, the name of the ACS journal, your **complete mailing address**, your 24-hour fax number, and the date that you need to receive our reply. For manuscripts in ASAP status, please use the RightsLink permission system to obtain permission.

*Publication reported in Section 5.1:* Casey A. Cole, Bethany Janos, Dien Anshari, James F. Thrasher, Scott Strayer, Homayoun Valafar. 2016. Proceedings of the International Conference on Health Informatics and Medical Systems (HIMS). Reprinted here with permission from the publisher (Springer Nature). The following shows Springer Nature's policy for reprint in Theses:

your licence via the email attached to your RightsLink receipt;

- Accept the terms and conditions and you're done!

For questions about using the RightsLink service, please contact Customer Support at Copyright Clearance Center via phone +1-855-239-3415 or +1-978-646-2777 or email [springernaturesupport@copyright.com](mailto:springernaturesupport@copyright.com).

### How to obtain permission to reuse Springer Nature content not available online on SpringerLink

Requests for permission to reuse content (e.g. figure or table, abstract, text excerpts) from Springer Nature publications currently not available online must be submitted in writing. Please be as detailed and specific as possible about what, where, how much, and why you wish to reuse the content.

#### Your contacts to obtain permission for the reuse of material from:

- books: [bookpermissions@springernature.com](mailto:bookpermissions@springernature.com)
- journals: [journalpermissions@springernature.com](mailto:journalpermissions@springernature.com)

### Author reuse

Please check the Copyright Transfer Statement (CTS) or Licence to Publish (LTP) that you have signed with Springer Nature to find further information about the reuse of your content.

Authors have the right to reuse their article's Version of Record, in whole or in part, in their own thesis. Additionally, they may reproduce and make available their thesis, including Springer Nature content, as required by their awarding academic institution. Authors must properly cite the published article in their thesis according to current citation standards.

Material from: 'AUTHOR, TITLE, JOURNAL TITLE, published [YEAR], [publisher - as it appears on our copyright page]'

If you are any doubt about whether your intended re-use is covered, please contact [journalpermissions@springernature.com](mailto:journalpermissions@springernature.com) for confirmation.

*Publication reported in Sections 5.2/5.3: Both submitted/accepted by JMIR. Reprinted here with permission from publisher. Please see answer provided below:*

[support.jmir.org/hc/en-us/articles/115001313067-Can-you-give-me-permission-to-publish-my-article-as-part-of-my-thesis-or-book-](https://support.jmir.org/hc/en-us/articles/115001313067-Can-you-give-me-permission-to-publish-my-article-as-part-of-my-thesis-or-book-)

[JMIR Publications](#) > [Production and Post-Publication](#) > [General Post-Publication Questions](#)

Search

#### Articles in this section

How to add an author after publication

Can I order reprints of my article?

Is JMIR Cancer in PubMed / PubMed Central yet?

Is JMIR Rehabilitation & Assistive Technologies in PubMed / PubMed Central yet?

Why does my article not have a PMCID yet?

Can you give me permission to publish my article as part of my thesis or book?

## Can you give me permission to publish my article as part of my thesis or book?



Editorial Director


3 months ago · Updated

Follow

Authors of articles published by JMIR Publications do not need our permission to use part or all of an article in their thesis or book. JMIR Publications' articles are published under a Creative Commons Attribution license (see copyright statement on each article), meaning that authors retain copyright ([What is a "Creative Commons License"?](#)). Authors can publish their article in a thesis, or as a book chapter, or anywhere else, as long as they adhere the requirements of the Creative Commons Attribution license, which is basically to cite the original source and to state that it was published (and can be reproduced) under the terms of Creative Commons Attribution license 2.0 (or 4.0 as of June 2017).

Authors should however be aware that -- while they retain copyright and may be able to use their work as book chapter, etc, as long as they disclose that it has been published as journal article in a JMIR journal --, [duplicate publication in other journals may be considered scientific misconduct](#).

*Publication reported in Section 5.4:* Casey A. Cole, James F. Thrasher, Scott Strayer, Homayoun Valafar. 2017. Contributed paper at 2017 IEEE International Conference on Biomedical and Health Informatics. Reprinted here with permission from the publisher as shown on the following website maintained by IEEE. It states that the work can be reposted anywhere without permission as long as the reader is directed to the IEEE copyright policy page: <https://www.ieee.org/publications/rights/copyright-policy.html>.



computer.org/conferences/cps/author-faq

TER

Conferences ▾ Publications ▾ Digital Library ▾ Education ▾ Membership ▾ Communities ▾ Get Involved ▾ Q

^ How do I submit my final paper PDF file?

▽ Who owns the paper's copyright?

If the conference content is under IEEE copyright, it is required that authors or their employers transfer copyright to IEEE, except for papers in the public domain or those that are reprinted with permission from a previously published, copyrighted publication. After transferring copyright to IEEE, authors and/or their companies may post their IEEE-copyrighted material on their websites without permission, provided that the sites alert readers to their obligations with respect to copyrighted material and that the posted work includes an IEEE copyright notice. [Please refer to the IEEE Copyright Policy page for more information.](#)

 ACS Publications  
Most Trusted. Most Cited. Most Read.

Search text, DOI, authors, etc. 

## Use in Theses/Dissertations

The following wording is from the ACS Thesis/Dissertation Policy and the ACS Journal Publishing Agreement:

**Reuse/Republishment of the Entire Work in Theses or Collections:** Authors may reuse all or part of the Submitted, Accepted or Published Work in a thesis or dissertation that the author writes and is required to submit to satisfy the criteria of degree-granting institutions. Such reuse is permitted subject to the ACS' ["Ethical Guidelines to Publication of Chemical Research"](#); the author should secure written confirmation (via letter or email) from the respective ACS journal editor(s) to avoid potential conflicts with journal prior publication\*/embargo policies. Appropriate citation of the Published Work must be made\*\*. If the thesis or dissertation to be published is in electronic format, a direct link to the Published Work must also be included using the [ACS Articles on Request](#) author-directed link.

\* Prior publication policies of ACS journals are posted on the [ACS website](#).

\*\* "Reprinted with permission from [COMPLETE REFERENCE CITATION]. Copyright [YEAR] American Chemical Society." Insert the appropriate wording in place of the capitalized words. This credit line wording should appear **on the first page of your ACS journal article**.

If your university requires written permission and your manuscript has not yet received a DOI (published ASAP), send a request to [copyright@acs.org](mailto:copyright@acs.org) that includes the manuscript number, the name of the ACS journal, your **complete mailing address**, your 24-hour fax number, and the date that you need to receive our reply. For manuscripts in ASAP status, please use the RightsLink permission system to obtain permission.