MACHINE LEARNING METHODS FOR SPECTRAL ANALYSIS

by

Youlin Liu

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Chemistry

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Garth Simpson, Chair
Department of Chemistry

Dr. Hilkka Kenttämaa
Department of Chemistry

Dr. Chengde Mao
Department of Chemistry

Dr. Chi Zhang
Department of Chemistry

Approved by:
Dr. Christine Hrycyna

To all the souls still trying to find their place: be patient

ACKNOWLEDGMENTS

This thesis would not have been possible were it not for the many, many wonderful people that I encountered. First and foremost, I would like to thank my parents for being there for me, and putting education as the most important pursuit for me always. It wasn't easy for a small-town girl born in rural China to be here today, my parents didn't understand scientific research or English at all for that matter, but they supported all of my decisions in searching for a self through higher education. I am lucky to have them.

My advisor, Garth, is an incredibly intelligent and hardworking researcher from whom I have learned not only how proper scientific research is conducted, but observed closely how going full-time academic researcher still means you are human, which is one of the most important things I've learned in my twenties. It's inspiring to watch him takes care of us while taking good care of his family. I really appreciate his quote, "Career, hobbies, family, you only need one of the three to not suck to live a happy life", which puts things into perspective.

Apart from Garth, James Ulcickas, and Casey Smith, two senior students in the Simpson Lab offered the most guidance, the contents that are not taught in classrooms, but needed in the Lab. They inspired me with their enthusiasm for work ethics.

I deeply appreciate Judy Liu and Anni Shi for their friendship as fellow grad students within the department. Grad school is like segmented spaces to a certain extend, in which the day-to-day contact is mostly people in the same lab. Judy is the one friend that I know would regularly reach out and check on me with the best of her intentions, and with Anni, we catch on departmental activities and fun stories. With them, I feel less alone in this sometimes struggling grad school experience.

Of course, there are many more people in the department that I very much appreciate. The Simpson Lab members, Alex Sherman, Andreas Geiger, who joined the lab the same year as me and I appreciate the shared experiences of hitting all of the milestones together. Changqin Ding, Zhengtian Song, Fengyuan Deng, and Shijie Zhang who are senior students who were also Chinese and took me in into the Chinese communities when I just arrived at Purdue; Hilary Florian, who I appreciate taking care of the logistics when we went to

Argonne national lab for the first time; Scott Griffin and Nita Takanti showed their support to me outside of the lab by attending my open-mic night at a local bar; And the rest of Simpson lab members, I thank all of you as well. I'm also grateful for Rob Reason, who would lend me his guitar and let me use the mailroom for brief guitar practice during the day, who also encouraged me to start singing in local bars.

Last but not least, Zhiyang Wang, who is the one true friend that offered emotional support, and reciprocated vulnerability. It is a rare friendship that has lasted since when we entered undergrad together, and one friend that I have the confidence that would last a lifetime. It's truly wonderful to know that I can always trust her, and be trusted. I wouldn't have made it through grad school without her emotional support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AI	artificial intelligence
ANN	artificial neural network
API	active pharmaceutical ingredients
CA	cluster analysis
DCM	dichloromethane
DLVO	theory named after Boris Derjaguin and Lev Landau, Evert Verwey and Theodoor Overbeek
DNN	deep neural network
FRAP	fluorescence recovery after photobleaching
GALDA	generative adversarial linear discriminant analysis
GAN	generative adversarial neural-network
IR	infrared
LDA	linear discriminate analysis
ML	machine learning
PAT	process analytical technology
PCA	principal component analysis
PDE	partial differential equations
PENN	physics encoded neural network
PINN	physics informed neural network
PMMA	polymethyl methacrylate
PLS-DA	partial least squares discriminate analysis
QCL	quantum cascade laser
ROC	receiver operating characteristic
SC	subcutaneous
SHG	second harmonic generation
TL	transfer learning
USAF	U.S. Air Force

ABSTRACT

Measurement science has seen fast growth of data in both volume and complexity in recent years, new algorithms and methodologies have been developed to aid the decision making in measurement sciences, and this process is automated for the liberation of labor. In light of the adversarial approaches shown in digital image processing, Chapter 2 demonstrate how the same attack is possible with spectroscopic data. Chapter 3 takes the question presented in Chapter 2 and optimized the classifier through an iterative approach. The optimized LDA was cross-validated and compared with other standard chemometrics methods, the application was extended to bi-distribution mineral Raman data. Chapter 4 focused on a novel Artificial Neural Network structure design with diffusion measurements; the architecture was tested both with simulated dataset and experimental dataset. Chapter 5 presents the construction of a novel infrared hyperspectral microscope for complex chemical compound classification, with detailed discussion in the segmentation of the images and choice of a classifier to choose.

1. INTRODUCTION

This dissertation contains four chapters of different projects, while the application space for each of these vary slightly, they cover the same overarching theme of using ML and algorithms to better understand, interpret and make use of the analytical chemistry data. Hence this introduction chapter will start by introducing the major advances in the community of computer sciences.

1.1 Recent Advances in Machine Learning and the Relevancy to Chemistry

Big data, machine learning, artificial intelligence, experienced an explosive amount of attention since the beginning of this century, While these terms are overly used to the extent of abuse, it's useful to step back and look at the practical definitions of each of these. AI focuses on using a non-human system to make optimal decisions, while ML is a subcategory of AI, covering the specific area of using algorithms that are computer-centered. Big data, is the application space, simply put it means "large volume of data"', the implied context, that differentiate "big data" from "not big data", is not only the volume, but also the complexity, including the collection sources, data structure and the velocity. The combination of these – aiming at eliminating humans – has done its magic in digital imaging processing, natural language processing, autonomous driving etc.

An example of major relevancy to this thesis (Chapter 3) is the advent of GANs (generative adversarial neural networks). The basic concept of GAN – training two classifiers to compete against each other – was brought around by Goodfellow *et al.* at the year of 2014, in other words, very recent. In the short 6 years followed, multiple models and variants have become widely adapted structures in applications concerning digital image processing, including adversarial autocoder[2], couple GANs[3], CycleGAN[4] deep convolutional GANs[5], to list just a few. The primary application for all of these is in digital image processing.

Image data fits perfectly in the context of "big data" in the recent digital world, on the other hand, data generated in chemistry, specifically analytical chemistry, remained relatively simple, the major limiting factor has long been instrumentation. From the readout of a simple balance or burette, to the more modern and developed mass spectrometer, optical spectrometer, and chromatography instruments, the analysis of such data was never complicated, more or less empirical, at most calls for statistical significance analysis. Needless to say, the data interpretation method development has gone stale in the recent years in the relevant communities: academic researchers, industries (The latter had a perfect reason to, due to limitations installed by FDA).

The most convenient adaptation, for application ML methods in chemistry, is to code the chemical information into vectors and matrices, then establish models attempting to predict how they would behave. Examples are chemical graph theory (correlation of chemical structures to biological activities [6]); chemical descriptors (similar to chemical graph conceptually, but via extracting numerical features from chemical structures [7]); chemical fingerprints (constructing high-dimensional vectors for virtual drug screening [8], or similarity analysis with exiting drug [9]); and there are more sporadic instances of using ML tools as is, for problems that can be described as "regression analysis" or "classification tasks" [10].

To this end, ML methods have been mainly been used in the science exploration space. In the meantime, downstream in the industry, a separate practice was to develop and implement innovative methods to pharmaceutical manufacturing processes, the FDA published guidance for the specifics, PAT (process analytical technology), to help regulate the development and implementation. Still, there exists a huge gap between the front-end computer science (that are actively being applied to digital companies such as Google Inc.) and how the pharmaceutical industry analyzes data, where the methods are data-driven and evaluated rigorously before final applications. The motivation of this thesis, is to explore the spaces in between.

1.2 Spectroscopic Data Analysis

Chapter 2 and 3 take a data-driven approach, and explored the analysis method with Raman spectra acquired for a drug with multiple crystal polymorphs. Therefore it's useful to step back at how spectroscopic data are generally handled, Raman of IR alike, The traditional process of spectroscopic data analysis, is to obtain a sample of interest, acquire three to five spectrum, then taking the average spectrum and identify main peaks of importance, and compare that of spectra of known compounds in numerous databases. Stepping into recent years, the broadening of applications of Raman and IR spectroscopy to various types of samples, sometimes heterogeneous in distributions, calls for a more systematic and reliable data handling approach. To gain deep insights into the critical information Raman and IR has to offer, the state-of-the-art tools included in multivariate analysis are principal component analysis (PCA), linear discriminant analysis (LDA), cluster analysis (CA) and partial least squares regression (PLS), to list just a few. These methods can alternatively be categorized as unsupervised methods and supervised methods, depending on whether label information is available or not. The central task when using these methods, was to extract the main features that is of interest, and representative of the sample of interest.

The observed spectral data is comprised of signal and noise, PCA is frequently used to extract main features, i.e. signals. Sometimes, the signal can be considered to have a correlation between features, it would be useful to capture the underlying correlation. Therefore, PCA can also be viewed as a dimension reduction method, in which it searches for representing the original data with fewer variables. To translate the original data into fewer variables, PCA produces principal components (PCs) and the translation can be done by matrix production. The way PCA produces the PCs are through eigendecomposition. Alternatively, LDA can be used to perform the same task, i.e. translate original data into fewer variables. LDA differs from PCA in the way that the PCs (in LDA it's termed "loading vector") are calculated: in LDA, the ratio of the between-class variance and the within-class variance is maximized, The use of (or lack of) class information divides these two algorithms into supervised and unsupervised. Detailed mathematical dissection of the two algorithms can be found in Section 2.2.1, thorough comparison of the methods has also been conducted Martinez and Kak[11].

As entailed LDA is one the simplest to implement supervised methods for dimension reduction and classification. Albeit a "slight" problem (curse of dimension, actually it's not a slight problem, but Chapter 3 addresses this problem as well) but is it always reliable, though? Chapter 2 demonstrate how LDA can be fooled, and Chapter 3 proposes a method to improve LDA so that the trick described in Chapter 2 attack doesn't work anymore. At this point, LDA is no longer LDA, the process of producing the loading vectors has been optimized via a generative approach, hence it was named GALDA (generative adversarial linear discriminant analysis).

1.3 Protein Diffusion Measurements

Proteins have emerged as a major class of therapeutics in recent years, and the subcutaneous (SC) injection of proteins is an appealing future as it's envisioning patients doing at-home treatments, making it much cheaper than doctor visits. However, it's the process after SC until the protein reaches the bloodstream crucially affects the bioavailability of the drug. Therefore it's useful for modeling the diffusion process, in a protein-matrix context. *In vitro* assessments of such interaction would provide information for the *in vivo* protein diffusion process as well.

The subcutaneous environment is complicated in chemical composition for the *in vitro* experiments to precisely simulate[12]. The potential issues and factors include PH, temperature, isoelectric point shifts of the protein, etc. It's unlikely the *in vitro* assays would adequately capture the inherent complexity, not to mention, the heterogeneity of substance distributions. Nevertheless, it's possible to simplify and parameterize the system with analytical methods. Chapter 4 is an attempt to characterize the protein/matrix cross interactions using protein diffusion measurements. While there are many instruments for accessing this property, FRAP (fluorescence recovery after photobleaching) stands out as one that incorporates into a high-throughput streamline afforded by the 96-well plate, which is compatible with the robotics structure already in use in the industry. Chapter 4 correlates diffusion constant with experimental variables in a novel structure of an artificial neural network.

1.4 Hyperspectral IR Imaging

Hyperspectral IR imaging (synonymous of IR spectroscopic imaging) as a technique, appeared to have matured and commercialized long before the "big data" time materialize. And rightfully so, given the simplicity promised by Beer's law, the ease of access of IR illuminating sources, and the range of information provided by molecular absorption. Since the arrival of the information age, there has been a recent renaissance of this technology, both in hardware development afforded by the advent of QCLs (quantum cascade lasers) and in data interpretation as the spectral data qualifies as high-dimensional data – a term that intimidates the average statistician.

Now complex methodology has been derived for a general high dimensional data, for example, the health data recorded including many attributes (patient age, gender, genetic background, drug consumption, etc), which in turn pushed forward the development of machine learning methods, as these applications (public health, biologics, etc) serves as a solid motivation to drive the methodology development forward. In light of those developments, Chapter 5 investigated into using multiple ML methods for both better optimization and classification of the spectral data.

References

[1] I. Goodfellow, P.-A. Jean, M. Mehdi, X. Bing, W.-F. David, O. Sherjil, C. Aaron, B. Yoshua, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 3, no. January, pp. 2672–2680, 2014, ISSN: 10495258. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[2] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, *Adversarial autoencoders*, 2016. arXiv: 1511.05644 [cs.LG].

[3] M.-Y. Liu and O. Tuzel, *Coupled generative adversarial networks*, 2016. arXiv: 1606.07536 [cs.CV].

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, *Unpaired image-to-image translation using cycle-consistent adversarial networks*, 2020. arXiv: 1703.10593 [cs.CV].

[5] A. Radford, L. Metz, and S. Chintala, *Unsupervised representation learning with deep convolutional generative adversarial networks*, 2016. arXiv: 1511.06434 [cs.LG].

[6] D. Bonchev, *Chemical graph theory: introduction and fundamentals.* CRC Press, 1991, vol. 1.

[7] A. U. Khan *et al.*, "Descriptors and their selection methods in qsar analysis: Paradigm for drug design," *Drug discovery today*, vol. 21, no. 8, pp. 1291–1302, 2016.

[8] J. W. Raymond and P. Willett, “Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2d chemical structure databases,” *Journal of computer-aided molecular design*, vol. 16, no. 1, pp. 59–71, 2002.

[9] R. P. Sheridan and S. K. Kearsley, “Why do we need so many chemical similarity search methods?” *Drug discovery today*, vol. 7, no. 17, pp. 903–911, 2002.

[10] Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, “Machine learning in chemoinformatics and drug discovery,” *Drug discovery today*, vol. 23, no. 8, pp. 1538–1546, 2018.

[11] A. M. Martinez and A. C. Kak, “Pca versus lda,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[12] H. M. Kinnunen, V. Sharma, L. R. Contreras-Rojas, Y. Yu, C. Alleman, A. Sreedhara, S. Fischer, L. Khawli, S. T. Yohe, D. Bumbaca, *et al.*, “A novel in vitro method to model the fate of subcutaneously administered biopharmaceuticals and associated formulation components,” *Journal of Controlled Release*, vol. 214, pp. 94–102, 2015.

2. ADVERSARIAL SPECTROSCOPY

2.1 Introduction

Modern instrumentation can produce ever-increasing volumes of measurements, the collective analysis of which is routinely used to inform chemical decision-making[1]–[3] As the dimensionality of the data increases, the greater it becomes our collective reliance on algorithmic data analysis approaches for quantification, classification, and basic interpretation[4]. With this exponentially increasing access to and reliance on large-data measurements, assurance in the outcomes impacts the legal and regulatory decisions made based on dimension reduction approaches such as deep neural networks[5], including drug testing[6], DNA matching[7], [8], regulation of pharmaceutical manufacturing[9], voice/facial recognition[10], etc. Given the societal importance of the outcomes in such instances, the motivation is high for ensuring reliable classification in data-intensive decisions[11], [12].

Artificial neural networks (ANNs) are known to exhibit potential susceptibilities to adversarial manipulations. Examples of successful attacks on ANNs trained for image recognition include subtle image perturbations, single pixel attacks, and adversarial stickers, among others. Subtle image perturbations that are visually difficult to identify visually because the magnitude of the attack at each pixel is 0.7% the magnitude of the original pixel value, yet the sum of these small perturbations cause an image to be misclassified. Single pixel attacks drastically change the value of a single pixel, such a change is obvious upon inspection of the image[13]. However, it has been shown that a single pixel is enough to unequivocally mis-classify an image[14]. Another method of attack has a user place a specially designed sticker into the field of view. The addition of this patch to the image has proven successful in misclassifying images. These are just a few examples of adversarial manipulations on ANNs among a rich body of manuscripts.

In this context, it is interesting to consider whether analogous attacks could be performed on standard dimension reduction methods such as linear discriminant analysis (LDA) and principal component analysis (PCA) as relatively simple and mathematically tractable surrogates for attacks in ANNs. For example, potential vulnerabilities may arise as a consequence of "over-fitting", in which the size of the training data set is small relative to the intrinsic dimensionality of the underlying data set and/or the number of spectral channels. In such instances, the statistical assessments derived from classification can be biased from numerical interpretation of noise as features for discrimination between classes. In addition to over-fitting, errors can arise when classes of testing spectra are not adequately represented within the training data set. Furthermore, spurious effects such as cosmic rays or calibration artifacts can also produce mis-classifications for reasons that are not trivially obvious within the reduced dimensional space. In all these instances, scientific decisions based on data-mining may be subject to additional uncertainties not necessarily reflected fully by the statistics evaluated within the reduced dimensional space.

In addition to linear methods serving as simple models for understanding adversarial attacks, their development also has the potential to share some of the benefits derived from applications of adversarial attacks in ANNs. In this work, we explored the potential applications of adversarial attacks in simple, linear dimension reduction methods for spectroscopic analysis, and we demonstrated a method for launching a digital adversarial spectroscopic attack and the statistical implications of such attacks. In designing the attacks, additive perturbations affected classification in a reduced-dimensional space defined by LDA.

2.2 Methods

2.2.1 Linear Discriminant Analysis

LDA was performed by solving the equation $J_w = \frac{\boldsymbol{W}^T \boldsymbol{S_B} \boldsymbol{W}}{\boldsymbol{W}^T \boldsymbol{S_W} \boldsymbol{W}}$, where J is the Fisher linear discriminant function (directly related to the resolution between classes), $\boldsymbol{W}$ is a matrix of the optimal projection vectors to maximize resolution (i.e., the eigenvectors, $\boldsymbol{S_W^{-1}} \boldsymbol{S_B}$), $\boldsymbol{S_W}$ is the within-class variance matrix, and $\boldsymbol{S_B}$ is the between-class variance matrix. Briefly, let $\boldsymbol{X_\mathrm{i}} = \{x_1, x_2, \ldots, x_n\}$ represents n_i training samples of class i, with the total number of classes given by c. Given μ_i as the centroid of $\boldsymbol{X_\mathrm{i}}$ then the within-class scatter matrix $\boldsymbol{S_W}$ and between-class scatter matrix $\boldsymbol{S_B}$ are defined as $\boldsymbol{S_W} = \sum_{\mathrm{i}=1}^{c} \sum_{x \in \boldsymbol{X_\mathrm{i}}} (x - \mu_\mathrm{i})(x - \mu_\mathrm{i})^T$ and $\boldsymbol{S_B} = n_{tot}^{-1} \sum_{\mathrm{i}=1}^{c} n_\mathrm{i} (\mu_\mathrm{i} - \mu_{tot})(\mu_\mathrm{i} - \mu_{tot})^T$, in which μ_{tot} is the weighted mean spectrum of the entire data set containing n_{tot} spectra. The computational stability of the matrix inversion operation was performed using Gaussian elimination of the product, $\boldsymbol{S_W^{-1}S_B}$ with the MATLAB operation ”A/B” (rather than through explicit evaluation of $\boldsymbol{S_W^{-1}S_B}$.

2.2.2 Launching an Attack to Generated Decoy Spectra

The central objective of the adversarial attack is to identify the perturbation, δ, that optimally alters the classification of an initial spectrum (e.g., random seed) to a target group, subject to constraints imposed by a cost function. For an initial spectrum, x_s, the perturbed spectrum, x', is given by $x' = x_s + \delta$. The general strategy in the optimization of the attack perturbation δ is illustrated in Figure 2.1, which is intended to serve as a graphical depiction of an attack depicted in the reduced-dimensional space. In this work, these reduced dimensions represent the principal axes produced by linear discriminant or principal component analyses. Each wavelength channel in the original spectral space resulted in a “nudge” to collectively contribute to the position of the spectrum in the reduced dimensional space. While the primary spectral features (indicated by the thin black arrows) combined to dictate the general position within the reduced dimensional space, randomness within the noise (indicated by the short red arrows in 2.1 A and B produced a spread about that mean position.

Additional “nudges” by perturbation to each wavelength channel of the original spectrum can relocate the position of the initial spectrum in the reduced dimensional space to one significantly closer to the target, as illustrated in Figure 2.1C. The vector of deviations $\boldsymbol{d}$ from the initial sample spectrum, x_s, to the “target”, x_t, in the reduced dimensional space is given by the following expression.

$$\boldsymbol{d} = \boldsymbol{D} \cdot [(x_s + \delta) - x_t] \tag{2.1}$$

The matrix $\boldsymbol{D}$ is comprised of the set of eigenvectors that project the high-dimensional data to a lower-dimensional space (such as PCA or LDA). In the absence of other considerations, the optimal perturbation, $\hat{\delta}$, will be one that maximizes the probability that the perturbed spectrum will be classified as the target. The reduction in dimension associated with $\boldsymbol{D}$ is an under determined problem, with, in general, an infinite number of selections for

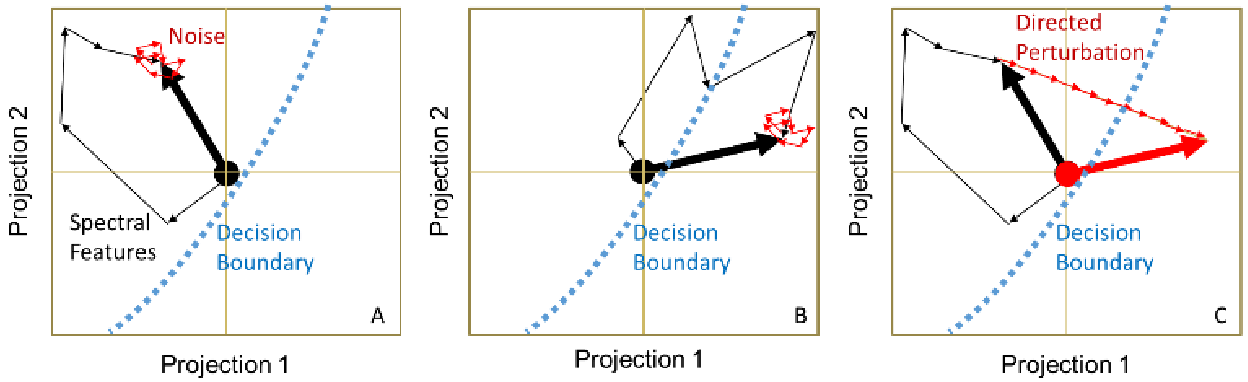


Figure 2.1. Conceptual illustration of spectroscopic adversarial attack. Major spectral peaks drive the position of spectra in lower-dimensional projections, demonstrated by the set of thin, long black arrows for Classes A and B, respectively. In C) the addition of patterned perturbations in the vector δ optimally relocates the position from the source Class A to Class B in this reduced dimensional space.

$\hat{\delta}$ producing comparable values for d. Therefore, $D \in R^{n \times p}$, with n being the dimensionality following dimension reduction.

The selection of one among the innumerable possible perturbations was performed by also considering an additional term in the cost function in Eq. 2.1 to minimize the squared magnitude of the perturbation. The total cost function for the reduced dimensional analysis was given by the sum of the two terms, which collectively minimized the sum of squared deviations to the target in the reduced dimensional space while simultaneously minimizing the overall squared magnitude of the perturbation, d, in spectral space, both evaluated as the squared L_2 norms.

$$\hat{\delta} = \arg\min_{\delta} \left[\left\| \boldsymbol{D} \cdot (x_s + \delta - x_t) \right\|_2^2 + \beta \left\| \delta \right\|_2^2 \right] \tag{2.2}$$

The scalar parameter β in Eq. 2.2 allows for empirical adjustment of the relative cost given to proximity to the target relative to minimizing perturbation of the major spectral features. The first term in the cost-function is designed to “fool the classifier” by minimizing the distance to the target in the reduced dimensional space, while the second term is targeted to “fool the human” by minimizing the perturbation in spectral space. In the present study, a value of β =1 was used throughout for simplicity.

The optimal perturbation was determined analytically by rewriting the cost function in terms of the perturbation.

$$F\left(\delta\right) = \left\| \boldsymbol{D} \cdot (x_s + \delta - x_t) \right\|_2^2 + \beta \left\| \delta \right\|_2^2 \tag{2.3}$$

To solve for the optimal perturbation, the minimum of $F\left(\delta\right)$ was found by setting the gradient to zero ($\nabla F = \mathbf{0}$). Let $C \;=\; \boldsymbol{D}^T\boldsymbol{D}$ and $v = x_s - x_t$:

$$\begin{aligned} F\left(\delta\right) &= \left\| C \cdot \left(v + \delta\right) \right\|_2^2 + \beta \left\| \delta \right\|_2^2 \\ &= v^T Cv + 2v^T C\delta + \delta^T (C + \beta I)\delta \end{aligned} \tag{2.4}$$

The following general gradient vector operations can be used to simplify Eq. 2.4:

$$f\left(x\right)=C^{T}x\Rightarrow\nabla f\left(x\right)=C \tag{2.5}$$

$$f\left(x\right)=x^{T}Ax\Rightarrow\nabla f\left(x\right)=\left(A^{T}+A\right)x \tag{2.6}$$

Using the relations in Eq. 2.5 and Eq. 2.6 noting that $C^{T}=C$, ∇F can be written in the following form.

$$\nabla F\left(\delta\right)=2\left(C+\beta I\right)\delta+2Cv \tag{2.7}$$

Setting $\nabla F = 0$ and solving for δ yields the following expression for the optimized perturbation, bearing in mind $C^{T}=C$.

$$\delta=-\left(C+\beta I\right)^{-1}C(x_{s}-x_{t}) \tag{2.8}$$

2.2.3 Probability

A two-class univariate problem, assuming normally, identically, and independently distributed measurements, is illustrated in Figure 2.2.

Probability is conveniently assessed using the z-statistic based on the normalized distances between the test value x_0 and the means μ of and standard deviations σ for each of the two distributions, defined as $z_1=\frac{x_0-\mu_1}{\sigma_1}$ and $z_2=\frac{x_0-\mu_2}{\sigma_2}$. The ratio r of probabilities P for a given scalar value of x_1 is given by:

$$r_{12}\equiv\frac{P_1}{P_2}=\frac{f(z_1)}{f(z_2)}=\frac{\mathrm{e}^{\frac{-{z_1}^2}{2}}}{\mathrm{e}^{\frac{-{z_2}^2}{2}}}=\mathrm{e}^{-\frac{1}{2}({z_1}^2-{z_2}^2)} \tag{2.9}$$

Given that $P_1+P_2=1$ and $\frac{P_1}{P_2}=r_{12}$, the probability for the test value x_1 corresponding to class 1 is given by $P_1=\frac{r_{12}}{(1+r_{12})}$, with $P_2=1-P_1$. The same argument can be made at every value in the reduced-dimensional spectrum x_0. Using the mean and standard deviation

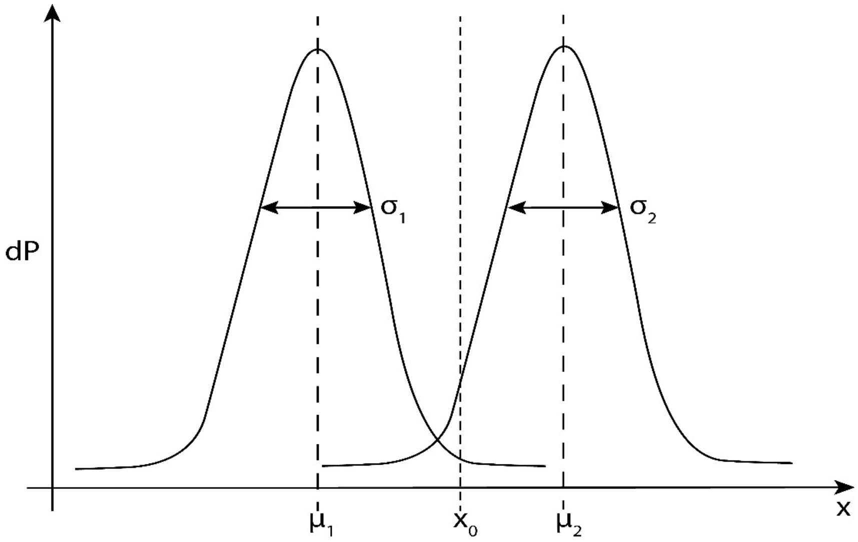


Figure 2.2. Illustration of the approach used for determination of the probability of a perturbed spectrum in the reduced dimensional space x_o to be classified as class 1 or class 2 (illustrated for a one-dimensional, two class system for illustrative purposes).

for each class in the reduced dimensional space, the z-statistic for a particular class n can be evaluated as a vector.

$$z_n = \frac{x_0 - \mu_n}{\sigma_n} \tag{2.10}$$

The probability ratio r_{12} for the vector description is straightforward to evaluate.

$$r_{12} = \mathrm{e}^{-\frac{1}{2}(z_1{}^T \cdot{}_1 - z_2{}^T \cdot{}_2)} \tag{2.11}$$

For an N-class system, P_n is given by the following.

$$P_n = \frac{1}{\frac{1}{r_{n1}} + \frac{1}{r_{n2}} + \frac{1}{r_{n3}} + \frac{1}{r_{nN}}} = \frac{1}{r_{1n} + r_{2n} + r_{3n} + ... + r_{Nn}} = \frac{1}{\sum_{\mathrm{i}=1}^{N} r_{\mathrm{i}n}} \tag{2.12}$$

2.2.4 Raman Measurements

Pure clopidogrel bisulfate Form I and Form II were produced in-house at Dr. Reddy's Laboratories and were used as received. Both the Form I and Form II particles were spherical with similar particle size distributions (diameter: ∼ 25 µm). Raman spectra were acquired using a custom Raman microscope, built in-house and described in detail previously[15]. In brief, a continuous wave diode laser (Toptica, 785 nm wavelength) coupled into a Raman probe (InPhotonics, RPS785/24) was collimated by a fused silica lens, and directed through an X-Y scan head composed of two galvanometer scanning mirrors. Two additional fused silica lenses formed a 4f configuration to deliver a collimated beam on the back of a 10x objective (Nikon). The Raman signal from the sample was collected through the same objective and de-scanned back through the same beam path into the Raman probe. A notch filter was built in the Raman probe to reject the laser signal. Raman spectra were acquired using an Acton SP-300i spectrometer with a 100 × 1340 CCD array, and controlled by a computer running WinSpec32 software. The laser power measured at the sample was around 30 mW. The exposure time was 0.5 s per spectral frame. To achieve higher signal to noise ratio for high quality training data for classification, 30 consecutive frames were averaged for each spectrum acquired over a spot size of ∼ 2-3 µm diameter within the field of view. A Savitzky-Golay filter was applied to smooth the spectra, and a rolling ball filter was used to remove the fluorescence background. Finally, the spectra were normalized to their integrated intensities, i.e., the area under the curves. A subset of 252 Raman spectra were collected from the clopidogrel samples and separated into three classes (84 spectra per class) – Form I, Form II, and background (glass slide). The ground truth identity of these samples was known a priori.

2.3 Results and Discussion

2.3.1 Raman Spectra

The mean spectra, average of 84 measurements, for three classes are shown in Figure 2.3 B. The spectra corresponding to the background were assigned as Class 3 (top spectra, black). The spectra belonging to the two polymorphs of clopidogrel bisulfate were classified as Classes 1 and 2 (red, middle and blue, bottom trace respectively). Class 3 shared one major feature of note, that being a large rolling peak around 1280 cm^{-1}. Spectra collected for Classes 1 and 2 showed clearly notable differences distinguishable by the relative peak intensities of the major features at $\sim$ 1019 and 1030 cm^{-1}, along with numerous minor peaks present in one or the other of the two sets of spectra.

LDA was performed on the data set of Raman spectra, as shown in 2.3A. Linear discriminant analysis provided clear separation between the different spectral classes upon dimension reduction. In order to solve for the optimal projections on LDA space, the inverse of the within-class variance matrix was matrix-multiplied by the between-class variance matrix. Additional spectra were generated to enable direct matrix inversion via knowledge of the mean and standard deviation of each wavelength for each class. These additional spectra allowed for the eigenvectors to be calculated to project the Raman spectra onto the lower dimensional LDA space.

2.3.2 Loading Vectors

The loading vectors from LDA to project Raman spectra onto the lower-dimensional space are shown in Figure 2.4. Inspection of the loading vectors shows that some of the

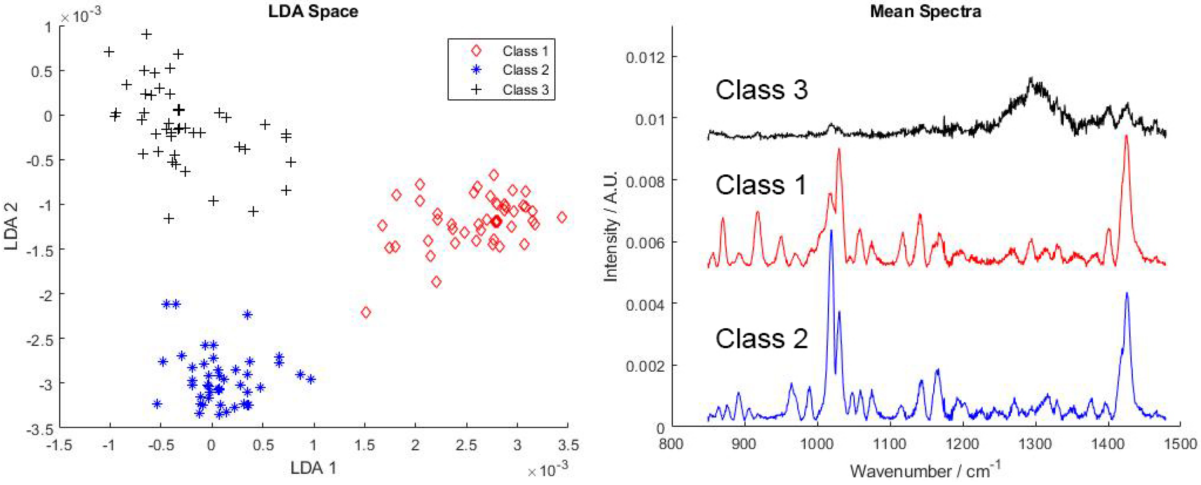


Figure 2.3. Projection of clopidogrel Raman spectra in LDA-space (on the left), with the corresponding mean spectrum for each class (on the right), offset for clarity. The mean spectrum for each class is an average of 84 repeated measurements.

spectral features of the Class 1 and 2 are present. Of particular note, however, is the high frequency content present in both eigenvectors. It will be shown later that the high frequency content is the main handle that the adversarial attack uses to induce mis-classifications. The underlying spectral features are present and support reliable classification in validation data sets, but are largely obscured by noise from the use of finite training data.

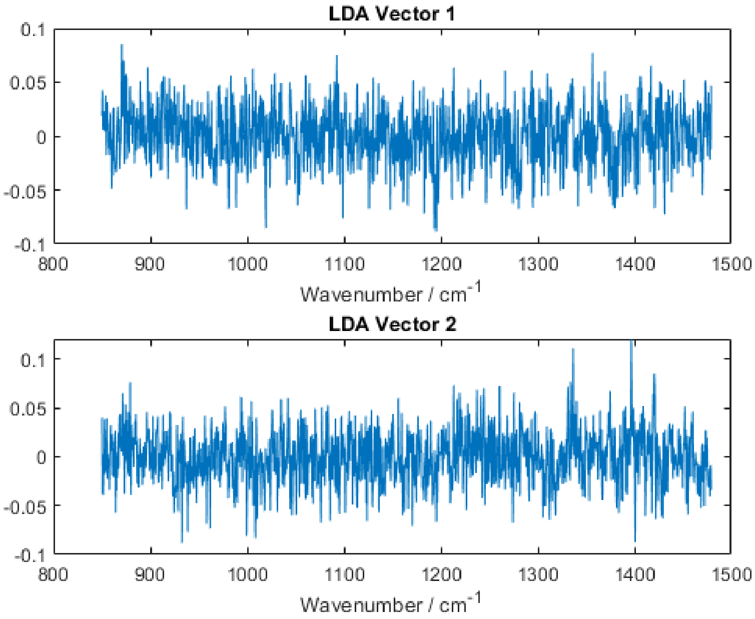


Figure 2.4. The loading vectors recovered from LDA to project Raman spectra onto the lower-dimensional LDA space.

2.3.3 Spectroscopic Adversarial Attack

The illustration of adversarial attack was first introduced by Goodfellow *et al.*[5] to support their hypothesis that neural networks are too linear to resit linear adversarial perturbations. Briefly, given an image, the classification of which by GoogLeNet can be purposely changed by adding an imperceptibly small vector (the perturbation), as shown in Figure 2.5.

The concept of spectroscopic adversarial attack is to translate the system from digital images to spectroscopic systems. As is shown in Figure 2.6. Upon visual inspection of the

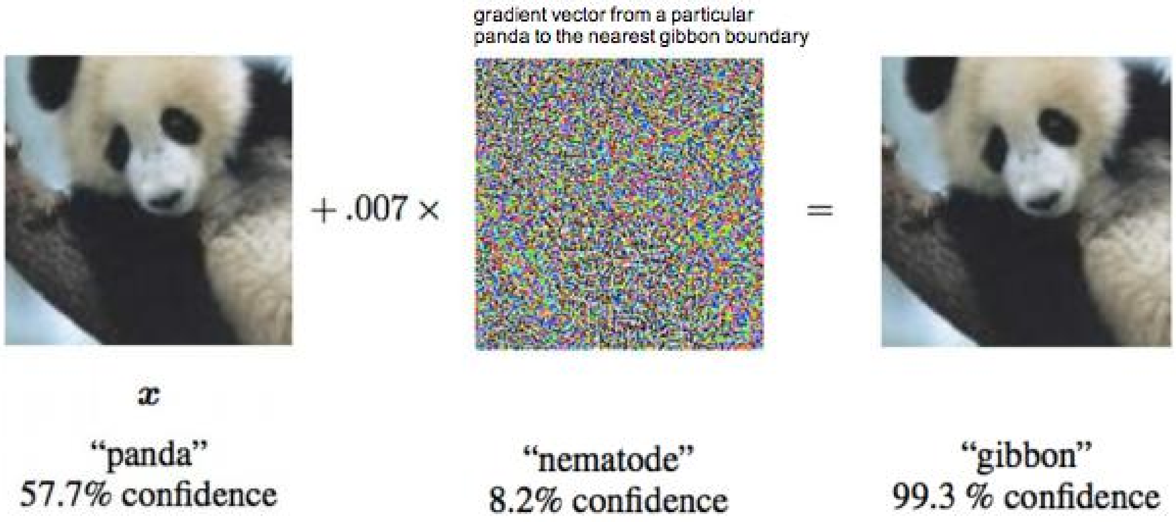
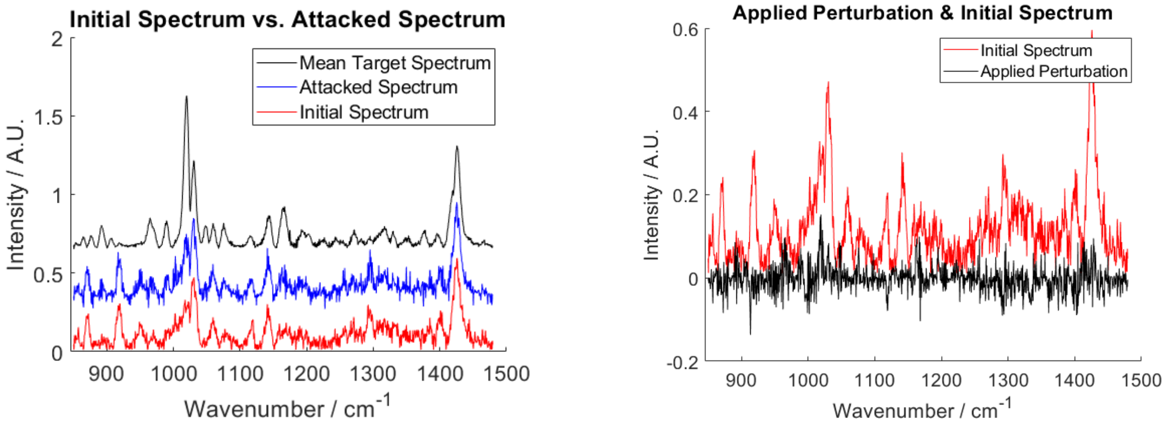


Figure 2.5. Demonstration of adversarial example by Goodfellow *et al.* in digital imaging processing. The perturbation was calculated by taking the elements of the gradient of the cost function with respect to the input. The classifier tested here is GoogLeNet[16].

perturbed spectrum (blue, middle) in comparison with the initial spectrum (red, bottom) in Figure 2.6a, no major spectral features can be identified as altered. The reason for which is that the calculated perturbation is of low amplitude as shown in 2.6b. It's interesting to note that the applied perturbation that optimally produce changes in classification are not due to changes in the major spectral features, as one might initially anticipate.

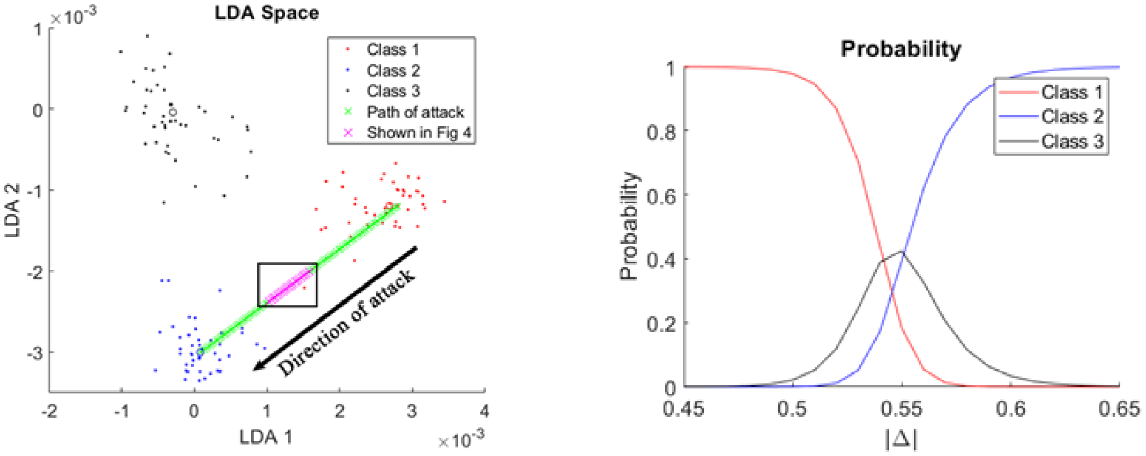


(a)

(b)

Figure 2.6. Spectroscopic adversarial attack illustration. 2.6a Comparison of the initial spectrum (red, bottom) from Class 1 with the attacked spectrum (blue, middle) classified as Class 2, and the mean target spectrum (black, top) of Class 2. 2.6b Comparison of the initial spectrum (red, top) and the applied perturbation (black, bottom). No offset applied.

Visualization of a representative spectroscopic adversarial attack in the reduced-dimensional LDA space is shown in Figure 2.7. Each perturbation was designed to displace the initial spectrum toward the target class within the reduced dimensional space. Because of the reduction in dimension, the direction of perturbation δ is under-determined, such that the optimal perturbation was calculated based on the cost function given by Eq. 2.1. The attack in Figure 2.7a was designed to move a spectrum from Class 1 (middle right, red) to Class 2 (bottom left, blue). The green x's represent the optimized perturbation, along the path from initial to target classification represented by $|\Delta|$, where $|\Delta| = 0$ corresponds to the initial, unperturbed spectrum and $|\Delta| = 1$ corresponds to the perturbed spectrum positioned at the mean of the target class in LDA-space. The purple x's represent the region of greatest uncertainty in classification, which is defined as the classification transitions from 95% confidence as the initial class to 95% confidence as the target class. This region of interest is highlighted in 2.7a. The full details for probability determination can be found in Section 2.2.3.



(a)

(b)

Figure 2.7. Adversarial attack in the LDA space. 2.7a: Demonstration of an incremental attack from Class 1 (red) to Class 2 (blue). Shown by the x's progressively moving towards the mean of Class 2, indicated by hollow blue circle. The probability of the at-tack belonging to each class was calculated at each point, x. The probability of the purple x's in the box belonging to each class is shown in Figure 2.7b. The purple x's denote the region of greatest uncertainty. 2.7b: Zoom in of the probability of the perturbation belonging to each class in the region of greatest uncertainty, as shown by the boxed, purple x's in Figure 2.7a. $|\Delta|$ represents the degree of displacement toward the mean of the target class.

2.3.4 Generalizability of the Spectroscopic Adversarial Attack

To confirmed the generalizability of performing spectroscopic adversarial attacks, an additional demonstration is performed in which the class assignments of the initial and attacked were changed. Previous results have been successfully replicated as shown in Figure

2.8. It is worth noting that, among the classes (clopidogrel bisulfate Form I, Form II, and background (glass slide) corresponding to class 1, 2 and 3), class 1 and 2 bear similar spectral features as they represent the polymorphs of the same chemical; class 3 being the background therefore bears no significant similar spectral feature to class 1 and 2. In this demonstration, class 3 was selected as the initial spectrum, it was successfully attacked regardless of features in the spectral space. This result is significant in that it confirms the theory that it was the high frequency contents that was driving the mis-classification across the decision boundary.

To assess the degree to which the adversarial results shown previously may be biased by random sampling, the process was repeated many times while only altering the initial and targeted spectrum class assignments. The general results shown in 2.9 suggest that the attacking strategy is robust in misclassifying any selected initial spectrum into desired targeted spectrum.

The relative magnitudes of the perturbations required to confidently induce misclassification are surprisingly small. The dominance of high frequency content in the perturbation highlights the significance of the variance of the signal in both the initial and target spectra, which appears to drive much of the spectral power in inducing misclassification. This outcome is somewhat surprising, as no constraints other than proximity to target and minimization of magnitude were imposed on the cost function given in Eq. 2.2. Analysis of the median magnitudes of the initial spectrum and the applied perturbation corresponding to $> 95\%$ confidence in misclassification show that the applied perturbation was $\sim$12% the magnitude of the initial spectrum, and it is sufficient to unequivocally alter the spectral classification. The graphical depiction in Figure 2.1 may provide some insights regarding the absence of similarity between the perturbation with either the initial or target spectra. From inspection of the figure, the most direct path from the initial to the target in the reduced dimensional space will generally not pass through the origin. As such, reductions of the major peaks in the initial spectrum and growth of the major peaks for the target spectrum will not generally correspond to the optimal perturbation.

These results highlight the growing challenges in ensuring statistical validity in regulatory, business, and legal decisions derived from data-intensive measurements that may be subject to incidental or intentional perturbation. As demonstrated herein, adversarial at-

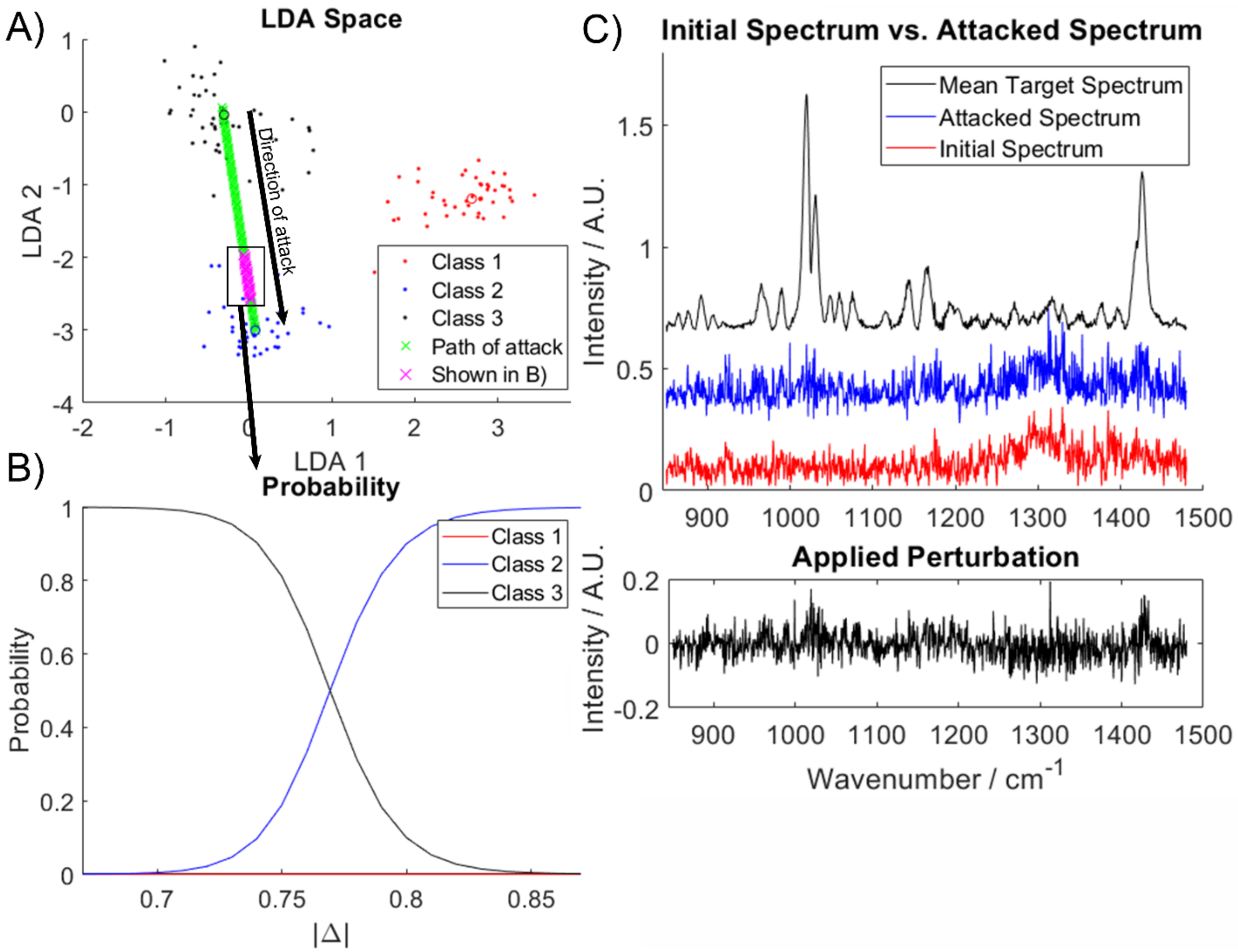


Figure 2.8. An additional demonstration of an "attack" on an initial spectrum from Class 3 to induce a misclassification as Class 2. A) The direction of the perturbation and the final position in LDA space are shown in the reduced dimensional space. B) The region of greatest uncertainty in classification is highlighted by the boxed, purple x's in panel A. $|\delta| = 0$ is the initial unperturbed spectrum, and $|\delta| = 1$ is the mean of class 2 in LDA space. At $|\delta| = 0.38$ there is a $> 95\%$ probability that the attacked spectrum belongs to Class 2 based on the position in LDA-space. C) Comparison of the initial spectrum (bottom, red) with the attacked spectrum at $|\delta| = 0.38$, which bears clear visual similarity to the initial unperturbed spectrum. D) Optimized perturbation used to induce misclassification.

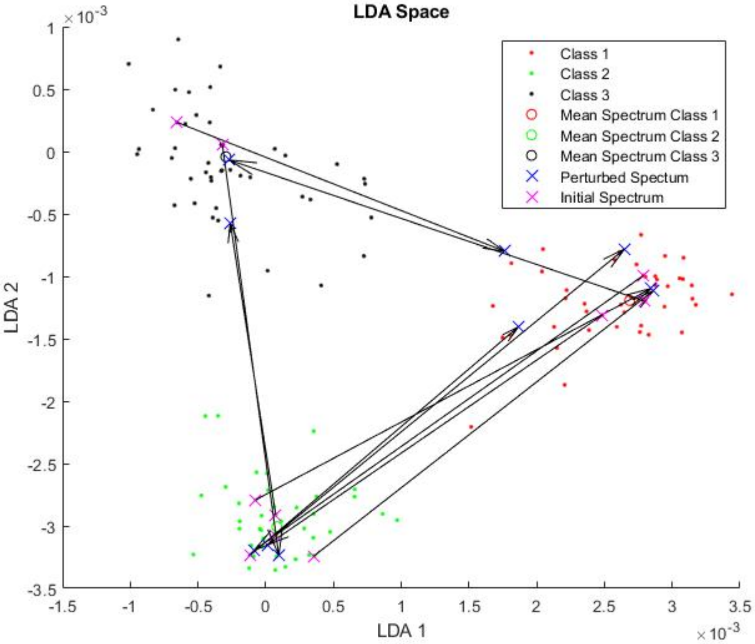


Figure 2.9. Attack strategies similar to those detailed in Figures 2.6 and 2.7 were performed for an additional 10 initial spectra, selected at random, for randomly selected target classification and spectra. All produced results were qualitatively similar to those detailed previously.

tacks on spectral information can profoundly influence decision outcomes, even when the perturbations are challenging to visually identify upon manual inspection. These results demonstrate that subtle digital alteration of files used in the otherwise benign operation of background subtraction can result in dramatically different outcomes in decision-making based on spectral analysis. Such relatively subtle changes spread over the entire spectrum would generally be challenging to discriminate from random noise. These results also demonstrate the vulnerability of dimension reduction techniques to instrument drift and cosmic rays among other common detector noise sources. As the volume of data integrated for decision-making increases along with the corresponding degree of pre-processing required for data mining, chemical decision-making is only poised to be increasingly susceptible to manipulation through adversarial perturbations. Attacks such as those demonstrated herein may also lead to future studies for improving chemical classifiers through generative adversarial strategies analogous. By identification of vulnerabilities, generative adversarial strategies may enable improvements in reliability and stability in conventional linear spectral analysis pipelines.

2.4 Conclusions

An optimized perturbation has been demonstrated to induce misclassification of an initial spectrum to a target class in a reduced-dimensional space. Interestingly, optimized perturbations did not contain obvious spectral features associated with either the initial or target class spectra, but rather appeared as noisy, featureless traces. Analysis of the median magnitudes of the initial spectrum and the applied perturbation corresponding to $> 95\%$ confidence in misclassification show that the applied perturbation was $\sim$ 12% the magnitude of the initial spectrum. These results highlight the hidden importance of residual high frequency content in defining the selection of principal coordinates for dimension reduction. Most significantly, awareness of this implicit sensitivity to unexpected high frequency spectral features in dictating classification may provide routes for improving robustness of dimension reduction methods and spectral classifiers to both noisy data and intentional alternations.

References

[1] E. Brynjolfsson, L. M. Hitt, and H. H. Kim, “Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?” *International Conference on Information Systems 2011, ICIS 2011*, vol. 1, pp. 541–558, Apr. 2011, ISSN: 1556-5068. DOI: 10.2139/ssrn.1819486. [Online]. Available: http://www.ssrn.com/abstract=1819486%20https://papers.ssrn.com/abstract=1819486.

[2] F. Provost and T. Fawcett, “Data Science and its Relationship to Big Data and Data-Driven Decision Making,” *Big Data*, vol. 1, no. 1, pp. 51–59, Mar. 2013, ISSN: 2167647X. DOI: 10.1089/big.2013.1508. [Online]. Available: https://www.liebertpub.com/doi/abs/10.1089/big.2013.1508%20http://www.liebertpub.com/doi/10.1089/big.2013.1508.

[3] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, ISSN: 00368075. DOI: 10.1126/science.1127647. [Online]. Available: http://arxiv.org/abs/physics/0601055www.sciencemag.org/cgi/content/full/313/5786/502/DC1%20%7B%5C%%7D3CGo%20to.

[4] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, N. Kruschwitz, S. LaVelle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, “Big data, analytics and the path from insights to value,” *MIT sloan management review*, vol. 52, no. 2, pp. 21–32, 2011. [Online]. Available: http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/.

[5] I. Goodfellow, P.-A. Jean, M. Mehdi, X. Bing, W.-F. David, O. Sherjil, C. Aaron, B. Yoshua, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, vol. 3, no. January, pp. 2672–2680, 2014, ISSN: 10495258. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[6] V. K. Ojha, K. Jackowski, V. Snášel, and A. Abraham, “Dimensionality reduction and prediction of the protein macromolecule dissolution profile,” in *Proceedings of the Fifth International Conference on Innovations in Bio-Inspired Computing and Applications IBICA 2014*, Springer, 2014, pp. 301–310.

[7] R. Clarke, H. W. Ressom, A. T. Wang, J. H. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang, “The properties of high-dimensional data spaces: implications for exploring gene and protein expression data,” *Nature Reviews Cancer*, vol. 8, no. 1, pp. 37–49, 2008, ISSN: 1474-175X. DOI: 10.1038/nrc2294. [Online]. Available: %7B%5C%%7D3CGo%20to.

[8] F. Model, P. Adorjan, A. Olek, and C. Piepenbrock, “Feature selection for dna methylation based cancer classification,” *Bioinformatics*, vol. 17, no. suppl_1, S157–S164, 2001.

[9] G. Reich, “Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications,” *Adv. Drug Deliv. Rev.*, vol. 57, no. 8, pp. 1109–1143, 2005, ISSN: 0169-409X. DOI: 10.1016/J.ADDR.2005.01.020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169409X05000578.

[10] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, “Discriminative common vectors for face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 4–13, Jan. 2005, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2005.9. [Online]. Available: http://ieeexplore.ieee.org/document/1359747/.

[11] J. Neumann, S. Christoph, S. Gabriele, C. Schnörr, and G. Steidl, “Combined SVM-Based Feature Selection and Classification,” *Machine Learning*, vol. 61, no. 1-3, pp. 129–150, Nov. 2005, ISSN: 0885-6125. DOI: 10.1007/s10994-005-1505-9. [Online]. Available: http://link.springer.com/10.1007/s10994-005-1505-9.

[12] C. Liberati and P. Mariani, “Big data meet pharmaceutical industry: An application on social media data,” in *Classification,(Big) Data Analysis and Statistical Learning*, Springer, 2018, pp. 23–30.

[13] J. Su, D. V. Vargas, K. Sakurai, and S. Kouichi, “One pixel attack for fooling deep neural networks,” *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019, ISSN: 0022-538X. arXiv: 1710.08864. [Online]. Available: http://arxiv.org/abs/1710.08864.

[14] T. B. Brown, Man, D. Mané, A. Roy, M. Abadi, and J. Gilmer, “Adversarial Patch,” no. Nips, 2017. arXiv: 1712.09665. [Online]. Available: http://arxiv.org/abs/1712.09665.

[15] S. J. Zhang, Z. T. Song, G. M. Dilshan, P. Godaliyadda, D. H. Ye, A. U. Chowdhury, A. Sengupta, G. T. Buzzard, C. A. Bouman, G. J. Simpson, G. Godaliyadda, D. H. Ye, A. U. Chowdhury, A. Sengupta, G. T. Buzzard, C. A. Bouman, and G. J. Simpson, “Dynamic Sparse Sampling for Confocal Raman Microscopy,” *Anal. Chem*, vol. 90, no. 7, pp. 4461–4469, 2018. DOI: 10.1021/acs.analchem.7b04749. [Online]. Available: https://pubs.acs.org/sharingguidelines.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, 2015. DOI: 10.1109/CVPR.2015.7298594.

3. GENERATIVE ADVERSARIAL LINEAR DISCRIMINANT ANALYSIS

3.1 Introduction

The increasing volume and complexity of information from modern chemical instrumentation places growing importance on methods to aid in data-intensive visualization and analysis[1], [2]. Dimension reduction is a key arrow in the analysis quiver, in which raw measurements in a high dimensional space (e.g., spectral space) are reduced to a handful of manageable feature-dimensions[3]–[5]. Transformation to feature-space provides several key advantages: i) ease of visualization of inherent clustering veiled at high dimension, ii) signal to noise enhancement through suppression of directions in measurement space that contribute predominantly to noise, and iii) improved simplicity for statistical hypothesis testing at low dimension. The number of methods available for dimension reduction is many, ranging from classic linear methods such as principal component analysis (PCA) and linear discriminant analysis (LDA) through complex neural network architectures for navigating highly nonlinear interrelationships in the data[6], [7]. However, even in these nonlinear cases, complementary analysis using linear methods can provide useful benchmarking and enable estimates for quantitative analysis using well-developed statistical methods for hypothesis testing[8].

When class information is available, LDA identifies the dimensions that maximize the resolution between the projected data through maximization of the Fisher linear discriminant function J[9]. Based on these metrics, one might initially expect LDA to be always preferred over PCA when classification information is available (i.e., with supervised data), as PCA neglects the additional classification information in the analysis. However, in practice, LDA is often ill-posed and prone to computational instabilities[10]. Such cases arise when either the number of training spectra is small, the number of classes is small, or both[11], [12]. The reason for the failure of LDA in these limits is tied to a matrix inversion step in its evaluation[13]. The Fisher linear discriminant function J maximized in LDA is given by the ratio of the between-class variance divided by the within-class variance, evaluated following projection of the input spectra onto a given test vector $\boldsymbol{w}$ in spectral-space or matrix of test vectors $\boldsymbol{W}$[14]. The linear algebra manipulations to determine the optimal projections $\boldsymbol{w}*$ to maximize J involve identification of the eigenvectors of the product of two matrices: one given by the inverse of a matrix describing the within-class variance $\boldsymbol{S_W}$ and the other describing the between-class variance $\boldsymbol{S_B}$. Because of the matrix inversion operation, the eigenvectors, and eigenvalues of $\boldsymbol{S_W^{-1}S_B}$ may be incalculable or exhibit large uncertainties. Such numerical instabilities are most likely to arise when the dimensionality of the measurements p greatly exceeds the number of replicates n in each class (i.e., $p > n$)[15], which is often the case in both spectral and image analyses.

A number of regularization strategies has been proposed for addressing computational instabilities associated with the matrix inversion step with $p > n$, and the corresponding propensities for overfitting that arise as a consequence. Arguably, the simplest regularization approach is the “shrunken centroids” method proposed by Guo, Hastie, and Tibshirani[16] in which scaled addition of an identity matrix to S_W results in a mathematically stable matrix inversion operation. Friedman also proposed a regularized discriminant analysis integrating linear and quadratic discriminant analysis, combined with a weighted identity matrix[17]. While generally very successful in enabling full-dimension LDA of spectral data, regularization can be quite sensitive to both the method and the degree of regularization[18]. Beyond regularization, a host of alternative approaches are available for feature extraction to reduce the spectral dimensionality as a means of overcoming the limitations of direct LDA at full spectral dimension. The simplest method is to first perform PCA as an initial dimension reduction step, then performed form LDA within the lower-dimensional PCA-space. While generally quite successful, the information lost by dimension reduction prior to LDA cannot be recovered by subsequent operations, such that the method adopted for feature extraction can significantly impact the quality of the final outcomes. Partial least-squared discriminant analysis (PLSDA) is also widely used for spectral classification[19], as well as dimension reduction[20]. Fourier transforms and discrete wavelet transforms have also been implemented for reducing dimensionality in spectral space prior to LDA[21].

On tackling overfitting arisen from the instability of linear algorithms, a plethora of machine learning (ML) methods have been demonstrated for this purpose. The most direct way would be to regularize the matrix inversion process, therefore regularizing the subsequent overfitting consequence. A simple way of introducing regularizing was proposed by Guo, Hastie, and Tibshirani, in which a scaled identity matrix is added to $\boldsymbol{S_W}$ to make it mathematical stable for matrix inversion[22]. There are many other approaches for doing the regularization[23], [24], still, this approach can be quite sensitive during implementation. Transfer learning strategies, broadly defined, have also been implements with considerable success for spectral analysis in the $p > n$ regime. Calibration transfer methods have a long and successful history of leveraging large volumes of well-characterized measurements to inform on spectral analyses when few case-specific spectra are available[25]–[27]. In work using artificial neural networks for spectral analysis, Li *et al.* developed a deep transfer learning based approach for near infrared spectroscopy multi-manufacturer drug identification[28]. This method achieved higher classification accuracy and scalability in multi-variety and multi-manufacturer NIR spectroscopy compared with current popular methods, such as support vector machines (SVM), back propagation (BP), the use of auto-encoders (AEs) and extreme learning machines (ELM). Zhang *et al.*[29] proposed an approach using a transfer-learning model pre-trained on a standard Raman spectral database for the identification of Raman spectra of organic compounds that were not included in the database and with limited data. Zhu *et al.*[30] first demonstrated usefulness and effectiveness of GANs for classification of hyper-spectral images (HSIs), using training samples to fine-tune a discriminative CNN for image classification. However, the application of the paper was limited to remote sensing with a focus on classification rather than dimension reduction, with no chemical spectral analysis. Later, Yu *et al.*[31] demonstrated classification of pathogens by Raman spectroscopy combined with generative adversarial networks to analyze the most salient identification regions in the real spectrum. Much of this collective body of work centers on classification, with fewer options centered around dimension reduction to aid in visualization of high dimensional spectral data.

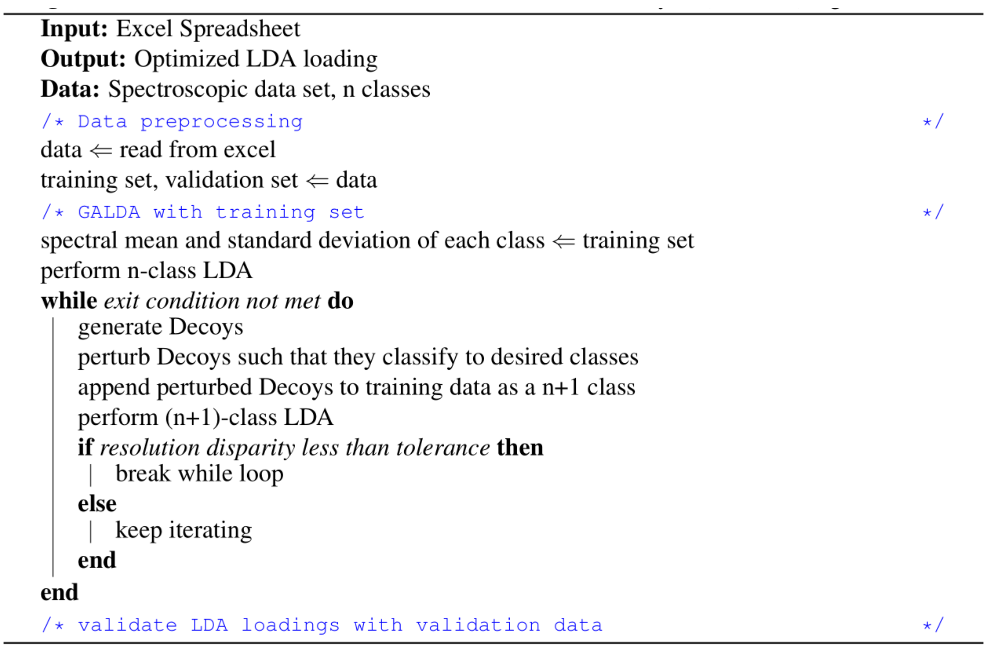
This numerical instability in full-dimensional LDA with small training size has some qualitative similarities to ”over-fitting” effects arising in machine learning tools, such as artificial neural networks[32]. Neural networks also generally possess a substantially greater number of adjustable parameters relative to the number of inputs used in their training[33]–[35]. In the limit of a small number of training spectra, repeated optimization of an ANN during training can result in increasing reliance on noise in driving classification/regression in addition to the signal. As a consequence, overfitting results in an increasing disparity between the accuracy of ANNs when evaluated with training versus testing datasets. In the case of ANNs, several strategies to address the consequences of overfitting in data-limited settings have found widespread adoption. The most common are arguably transfer learning (TL)[36] and the use of generative adversarial networks (GANs)[37]. By analogy with calibration transfer in chemometrics introduced in the preceding paragraph, TL in neural network applications leverages a pre-trained network to serve as a foundation for the extension to new systems[38]. In applications involving neural networks, TL requires access to a pre-trained network, which may not be available in many instances. One implementation of TL is through physics-informed simulations for data augmentation approaches introduced in the discussion of spectral data are also widely used in neural networks[39], in which neural networks are trained taking into consideration partial differential equations that represent physics-informed constrains. Another common approach is the use of generative adversarial networks (GANs), in which a second competing neural network or algorithm is introduced to improve the statistical reliability of an ANN. In brief, the generator is most often designed to convert a random initial seed to an input registered as being genuine and of a particular target class. A re-optimized ANN is produced to reject the generated "decoy" data as false, which is then targeted again by an updated GAN. Iteration between attack and defense improves the broader utility of the ANN in data-limited applications[40]. Between TL networks and GANs, GANs have the broader utility of the two approaches, as they can operate wholly independently for a given new problem without prior knowledge of related systems or pre-training.

In the previous chapter, we successfully demonstrated adversarial attacks on Raman spectral data, in which the perturbations were imperceivable upon manual inspection, but altered the classification of the given spectrum in the reduced space. This demonstration highlighted the vulnerability of dimension reduction techniques, which are common practices in the decision-making processes routinely done in the spectral analysis. In light of the successes of generative adversarial approaches to minimize over-fitting artifacts in ANNs[41], we hypothesize that analogous benefits may be realized to address the numerical instabilities complicating the use of LDA at full dimension in data-limited applications. To test this hypothesis, we developed an analog of the nonlinear processes intrinsic in GANs, but built around linear transformations inherent in LDA. Specifically, we developed a linear mathematical framework for optimally perturbing a random input seed to generate decoy spectra and used full dimension LDA to optimally separate genuine and generated training data, iteratively optimized to compete against each other. This proposed GALDA approach was compared with other common linear methods for dimension reduction using simulated spectra generated from an archived spectral database. Following these proof-of-concept studies, GALDA was used for dimension reduction of a set of spectra from Raman microscopy of clopidogrel bisulfate micro spheroids. Figures of merit for comparison across different methods for dimension reduction included: i) the resolution between classes in the reduced dimensional-space, ii) the degree of over-fitting assessed by the resolution difference between testing and training data, and iii) the smoothness/roughness of the corresponding "loading plots" onto which the raw spectra were projected for dimension reduction. These metrics were performed using the closely related methods of PCA, PLS-DA, PCA-LDA, and regularized LDA as comparators for GALDA. The strengths and limitations of GALDA were then critically evaluated in spectro-chemical analysis test cases for dimension reduction of classified data.

3.2 Methods

3.2.1 The Architecture of the GALDA Algorithm

In a typical GAN structure, two neural networks are set up such that they are competing against one another (thus "adversarial"). The two competing neural networks are a) a generative component and b) a discriminative component. In GALDA, the "generative" component is the generation of decoy spectra, which are the combination of "generate random noise" and "adversarially attack noise to become decoy spectra", the discriminative component is the (n + 1) class LDA. Figure 3.1 displays the pseudo code illustrating the logic of the GALDA algorithm. Briefly, the algorithm takes input structured as excel sheets, each of the sheet contains repeated measurements of the same class by the column; the data would then split into a training set and a validation set. Using the loading vectors produced via n-class LDA, randomly generated noisy spectra are perturbed into desired classes, also randomly selected. The combined process of n-class LDA and perturbation is the “generator”. Afterwards, a $n+1$-class LDA serves as the “discriminator”, separating the genuine data and the generated data, and updating the loading vectors. This process is iterated. The updated loading vectors at the end of each loop are used to calculate resolution for both the training and the validation data set. The iteration exits the loop when the resolution disparity between the two is less than a preset desired value.



```
Input: Excel Spreadsheet
Output: Optimized LDA loading
Data: Spectroscopic data set, n classes
/* Data preprocessing                                          */
data ⇐ read from excel
training set, validation set ⇐ data
/* GALDA with training set                                     */
spectral mean and standard deviation of each class ⇐ training set
perform n-class LDA
while exit condition not met do
    generate Decoys
    perturb Decoys such that they classify to desired classes
    append perturbed Decoys to training data as a n+1 class
    perform (n+1)-class LDA
    if resolution disparity less than tolerance then
        break while loop
    else
        keep iterating
    end
end
/* validate LDA loadings with validation data                  */
```

Figure 3.1. Pseudo code illustrating the logic of the GALDA algorithm. “resolution disparity” refers to the disparity between the training and validation dataset.

Figure 3.2 provides an overview of the work-flow for GALDA, in which LDA is used for dimension reduction. Classified initial input data in a) are projected into a reduced dimensional space defined by LDA, indicated in b). Next, a generative adversarial iteration (c through e) is performed. The panel in c) shows the projection of randomly generated inputs serving as initial decoy data projected into LDA-space. In d), the decoy data are then modified by the linear addition of a perturbation in spectral space to induce classification as one of the input classes within LDA-space. In e), LDA is performed again, but now with an additional class for the decoy data. The generation of decoy data, perturbations, and LDA is performed iteratively until convergence is achieved (c through e).

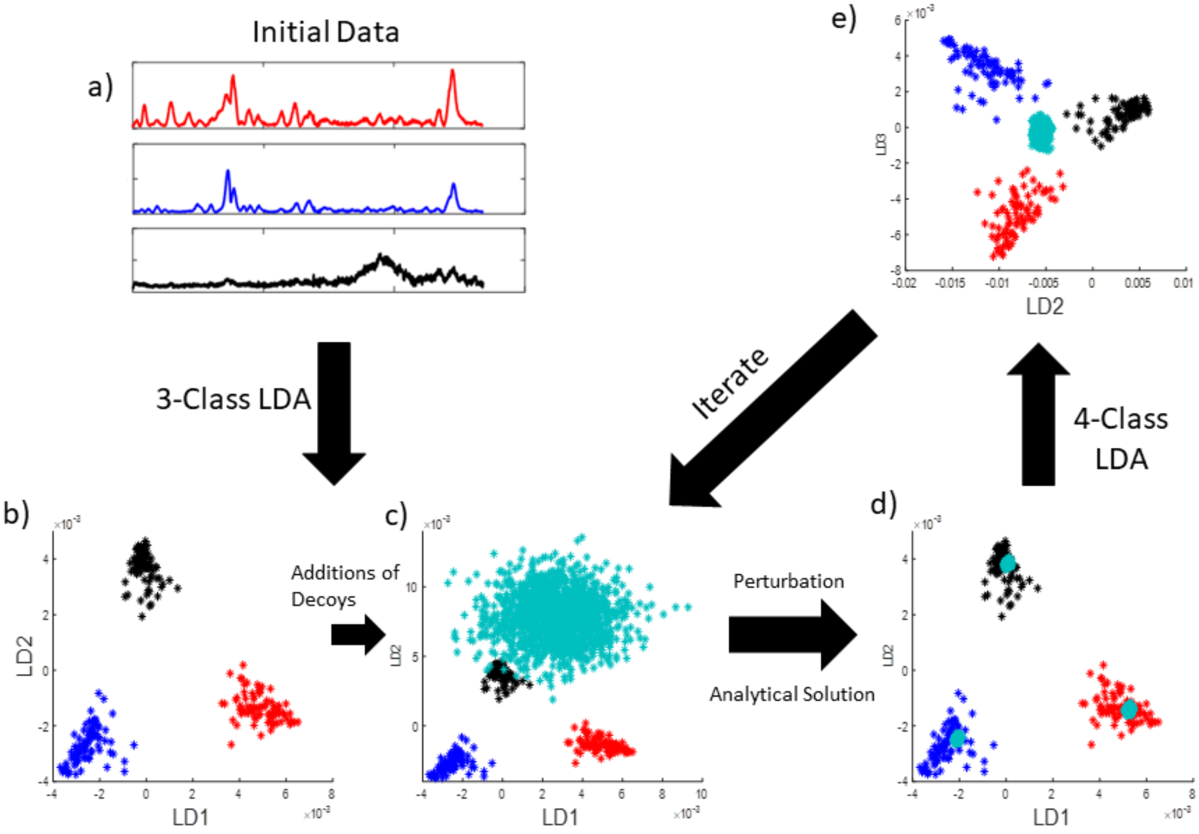


Figure 3.2. Work-flow of the generative adversarial linear analysis algorithm. Initial data that fall into three classes are projected onto a reduced dimensional space. Random inputs serve as initial decoy data projected onto this LDA-space, depicted here in 2D. The decoy data are then "attacked" to induce classification as one of the original input classes in LDA-space while minimizing the magnitude of the perturbation in spectral space. LDA is then performed again, but now with an additional class and corresponding dimension for isolating the decoy data. The processes c-e are iterated until convergence is achieved.

The script to perform gradient calculations is available at GALDA Public Repository along with the other scripts for performing adversarial and GALDA iterations.

3.2.2 Stopping Criteria

In this initial demonstration of GALDA, many key parameters were selected empirically and could be subject to further improvements in future implementations. As one example, the stopping criterion was based on empirical inspection of the trends in J_{tot} on a case-by-case basis rather than a formal mathematical assessment. Future implementations with automated stopping criteria are currently under consideration. The derivative of J_{tot} evaluated for the training data may serve as a viable metric, provided the assessment is performed over sufficient numbers of iterations to remove uncertainties in the slope from the "zig-zag" features in the low training-set limit. Alternatively, generalized assessments could be evaluated based on the size of each input (e.g., spectrum) and the number of inputs within each class. In addition to the stopping criterion, the number of spectra within each iteration of the generative loop has also not been optimized in this initial implementation. Optimization of these and other meta-parameters associated with the GALDA implementation may provide further improvements in achieved resolution and utility.

3.2.3 Resolution Definition

The eigenvalues J produced by LDA correspond directly to metrics of resolution between the different classes when projected along the corresponding LDA eigenvectors, serving as a convenient quantity for assessing the merits of different dimension reduction strategies. We introduce an algorithm to recover analogous values of J analytically that are applicable for any projection vector w irrespective of whether it is an eigenvector (e.g., for testing data, or for testing/training data using other dimension reduction methods). As expected, the analytical result described below recovers the eigenvalues from LDA when used with the corresponding eigenvectors. For a given selection of w, the dot product of w with a given spectrum x results in a scalar projection x. For a two-class (A,B) system with the populations projected onto a single coordinate, a unit-less form of the resolution R between the two classes of scalar-valued projections is given by $R = \frac{|\bar{x}_A - \bar{x}_B|}{\sqrt{s_A^2 + s_B^2}}$, where $\bar{x}_\mathrm{i}$ is the scalar mean of class i and s_i is the experimental standard deviation about the measured mean $\bar{x}_\mathrm{i}$ for class i. This definition of resolution holds for any particular selection of vectors w on which the data are projected, with some projections yielding higher or lower values for the resolution between the classes. In this two-class system, the Fisher discriminant J recovered by solving the equation $J_w = \frac{\boldsymbol{W}^T \boldsymbol{S_B W}}{\boldsymbol{W}^T \boldsymbol{S_W W}}$ from the eigenvectors/eigenvalues of $\boldsymbol{S_W}^{(-1)} \boldsymbol{S_B}$ yields the eigenvector $\boldsymbol{w}$ that maximizes J. Less obviously, the recovered value of J in a two-class system is identically equal to the squared resolution $J = R^2 = \frac{|\bar{x}_A - \bar{x}_B|^2}{s_A^2 + s_B^2}$, given by the between-class scalar variance divided by the combined within-class scalar variance following projection along $\boldsymbol{w}$. As such, maximizing J through solving the eigenvector/eigenvalue problem also corresponds to maximizing the resolution R. Generalization of the resolution to c-classes of scalar valued projections yields the following expression, calculable for any selection of $\boldsymbol{w}$ using the scalar projected values $x_\mathrm{i} = x^T \boldsymbol{w}$.

$$R^2 = J_{tot} = n_{tot}^{-1} \sum_\mathrm{i} n_\mathrm{i} \left(\overline{x}_\mathrm{i} - \overline{x}_{tot}\right)^2 / \left(\sum_\mathrm{i} s_\mathrm{i}^2\right) \tag{3.1}$$

In this equation, $\bar{x}_\mathrm{i} = \bar{x}^T \boldsymbol{w}$ is the experimental mean spectrum of class i, $\bar{x}_{tot} = \bar{x}_{tot}^T \boldsymbol{w}$ is the weighted mean spectrum of all c classes , σ_i^2 is the variance about $\bar{x}_\mathrm{i}$ for the projections of class i, n_i is the number of spectra within class i, n_{tot} is the total number of spectra in the training set, and the summations span $i = \{1 : c\}$. When $\boldsymbol{w}$ is an eigenvector of $\boldsymbol{S_W}^{-1} \boldsymbol{S_B}$, the value of J calculated analytically from projection and summation using Equation 3.1 is identically equal to the eigenvalue recovered for the Fisher linear discriminant J by LDA. However, Equation 3.1 has the advantage of also allowing the calculation of analogous resolution values when $\boldsymbol{w}$ is not an eigenvector, including evaluation of testing data. A value of J can be calculated for each selection of $\boldsymbol{w}$ if the reduced dimensional space has greater than one dimension, with $J_{tot} = \sum_\mathrm{j}^{c-1} J_\mathrm{j}$. This algorithm for evaluating J for 2-class data and J_{tot} for c-class data has the distinct advantage of enabling direct comparisons between the resolutions produced within the training data (which also correspond to the eigenvalues) and independent testing data not used in the training and therefore not amenable eigenvector/eigenvalue analyses. Further, this definition enables ”apples to apples” comparison of resolution metrics for other dimension reduction approaches.

3.2.4 Matrix Inversion

We elaborate here on the discussion of the ill-posed problem of matrix inversion inherent in LDA when the number of measurements n is less than the dimensionality in the spectrum p, and the corresponding impact on the iterative GALDA solution. Several computational approaches are available for approximating or evaluating the inverse, two of which were considered in the present study. First, it was noted that the matrix inversion step is one part of a matrix product, with the eigenvectors of $\boldsymbol{S_W^{-1}}\boldsymbol{S_B}$ corresponding to the reduced dimensions in LDA-space. Solving the combined product in Matlab via Gaussian elimination still results in an ill-posed problem for dimension LDA, with an infinite number of solutions to the under-determined problem. Consistent with this expectation, evaluation of full-dimension LDA produced a resolution within the training data approaching infinity and a resolution in the testing data approaching zero, as shown in the first few iterations of GALDA in Figure 3.5A with $p > n$. With 100 generated spectra per iteration, the combined number of real and generates spectra after 12 iterations was 1367 (1300 generated spectra combined with 67 genuine spectra). This is the first iteration for which the matrix inversion operation is no longer ill-posed for this dataset with spectra containing 1340 dimensions. From that point onward, the matrix inversion operation is equivalent by both Gaussian elimination and the Moore-Penrose pseudo-inverse. Consistent with these expectations, curves of testing and training resolution that deviated significantly when $p > n$ converge to similar outcomes in this asymptotic limit of $p << n$ with n evaluated using both genuine and decoy data.

3.2.5 Raman Simulations

Simulations were performed to assess the merits of the GALDA algorithm for a dataset with known ground-truth results. Six Raman spectra selected from the Romanian Database of Raman Spectroscopy[42] (magnetite, galena, molybdenite, goethite, stibnite, pyrolusite) recorded with 1024 wavelength channels served as ground truth source spectra. Following

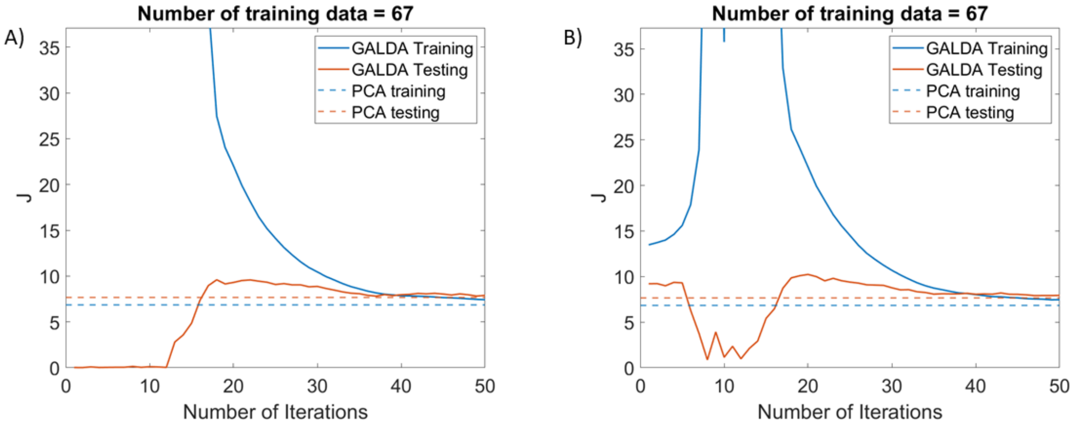


Figure 3.3. Evaluation of testing and training resolution for full-dimension LDA evaluated using Gaussian elimination to perform matrix inversion (A) verses via Moore-Penrose pseudo-inverse (B) in the iterative GALDA algorithm.

normalization, 100 simulated spectra for each of the six were generated by addition of Gaussian distributed noise with a mean of 0 and a standard deviation of 0.02. The spectra were grouped pairwise into three classes, resulting in a nontrivial, bimodal distribution within each class. This pairwise grouping (analogous to a single beach with two kinds of pebbles) was designed as a simple model for classes with nontrivial probability distributions, routinely encountered in practical spectral decomposition analyses.

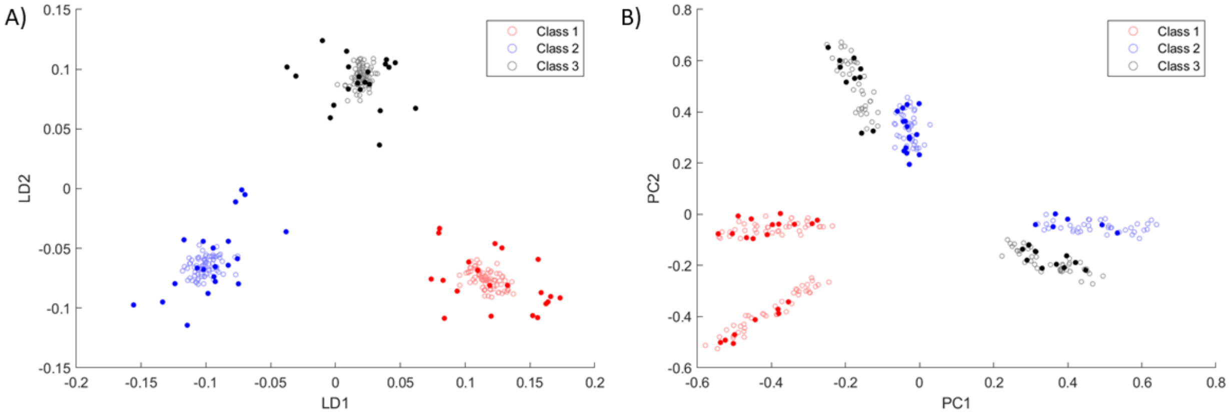


Figure 3.4. LDA and PCA dimension on simulated Raman spectra, with each class containing a binary mixture of mineral spectra with additive normally distributed noise: A) LDA, and B) PCA. The total number of spectra used in the analysis was 100 per class, split here into an 80/20 training versus testing dataset. Training data are indicated by open dots, testing data by filled dots.

A simple dimension reduction of the simulated data of bimodal distribution is visualized here using LDA and PCA for comparison. The simulations were designed to exemplify that PCA — as an unsupervised method — under-performs when data is of nontrivial distribution. Details of the simulations can be found in the Methods section in the main manuscript. Briefly, three classes of Raman spectra were simulated with 100 spectra per class. LDA and PCA were performed respectively as shown in Figure 3.4 A (LDA) and B (PCA). To make $\boldsymbol{S_W}$ and $\boldsymbol{S_B}$ square matrices (and invertible), additional spectra were generated via knowledge of the mean and standard deviation of each wavelength for each class.

Several key trends are apparent from inspection of the simulated spectra projected onto two-dimensional PCA and LDA-spaces. From Figure 3.4, differences in the variance about the centroid are clearly evident for the training data (open) and testing data (filled), consistent with significant overfitting. More strikingly, the PCA results illustrate the origin of the overall poor resolution between the three classes of data in PCA-space. Because each class consisted of a combination of spectra from two distinct populations, maximization of the total variance neglecting class information spread data from a single class across multiple regions within PCA-space. These results illustrate the susceptibility of PCA to poor resolution in analyses of data with nontrivial probability distributions within spectral space. However, PCA nevertheless produces good agreement between the distributions recovered for the testing and training datasets.

3.2.6 Raman Measurements

In addition to Raman simulations as mentioned above, experimental Raman data were also employed to test the GALDA algorithm. Details of the dataset can be found in the previous chapter (see 2.2.4).

3.3 Results and Discussion

3.3.1 Visualization of the Overfitting Problem

Before considering GALDA, it is useful to illustrate the impact of overfitting artifacts in the application of full dimension LDA in under-determined systems. Complications from overfitting in full dimension LDA are illustrated in Figure 3.5, using PCA as a comparator. Briefly, Raman spectra with 1340 wavelength channels were subject to dimension reduction by LDA (Figure 3.5A) and PCA (Figure 3.5B). To make $\boldsymbol{S_W}$ and $\boldsymbol{S_B}$ square matrices (and invertible), additional spectra were generated via knowledge of the mean and standard deviation of each wavelength for each class. A total of 84 Raman spectra were divided into training and testing data with an 80:20 split, respectively (67 spectra for training and 17 spectra for testing). Evidence for overfitting by LDA is shown in Figure 3.5A, in which a substantial disparity arises between the distributions of the testing and training data. Specifically, the training data were tightly distributed and not representative of the corresponding distribution in the testing data.

This result is an archetypal example of overfitting, as routinely arises with LDA and numerous other machine learning feature extractors in cases of low replicate numbers of high dimensional data[43]. In contrast, PCA is significantly less susceptible to overfitting under these same sample-limited conditions, as the distributions of both training and testing data for this data set are qualitatively similar. However, PCA is an unsupervised method, such that all the additional information associated with class label is not integrated into the dimension reduction. Consequently, the recovered eigenvectors from PCA are not generally optimized for class resolution. As a result, general qualitative guidance for selecting between LDA and PCA in supervised datasets recommends the use of PCA for small numbers of high dimensional data and LDA for low-dimensional data and/or with large numbers of classes[43]. The conditions corresponding to the tipping point between the two effects (loss in resolution from overfitting for LDA and from neglect of class information in PCA) are rarely immediately obvious and are highly data dependent.

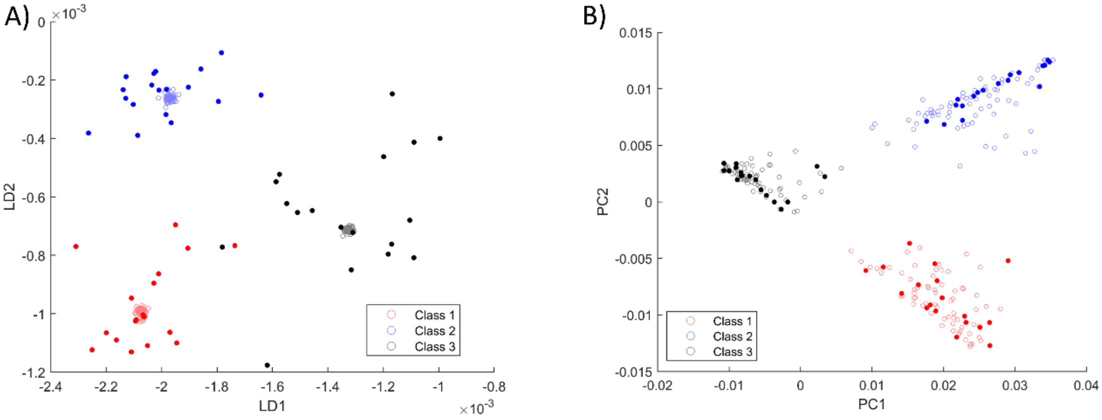


Figure 3.5. Illustration of overfitting in dimension reduction using LDA in (A) and PCA in (B). Classes 1, 2, and 3 correspond to clopidogrel bisulfate forms I and II, and background, respectively. The total number of spectra used in the analysis was 84 per class, split here into an 80/20 training versus testing dataset. Training data are indicated by open dots, testing data by filled dots.

3.3.2 The Iterations in GALDA Before Convergence

Figure 3.6 provides an overview of the "migration of location" in the reduced dimension during the GALDA iterative process for the Raman spectroscopy of clopidogrel bisulfate. Initially, a 3-class LDA of the training data (67 spectra from each class), resulted in a clear separation between the three classes (Form I, Form II, background), indicated by the red, blue, and black projected data points, respectively. Next, uniformly distributed random spectra (seed data) were generated and projected onto this initializing 2D LDA-space, resulting in a broad distribution indicated by the magenta data points in Figure 3.6A. Perturbation through adversarial attack in spectral space transformed the seed data to decoy data, which subsequently projected close to targets randomly assigned from the training set, as demonstrated in B. Then, performing a 4-class LDA forced the decoy data to separate from the training data by assigning the decoy data as an additional class.

Upon projection in 3D space, the LDA coordinate with the greatest eigenvalue (LD1) distinctly separates the training and decoy data. The remaining two coordinates (LD2 and LD3) serve as the updated reduced dimensional LDA-space for the training data. Using this "attack and defend" strategy, randomly generated seed data were perturbed to produce decoy data, which were separable along an additional dimension in LDA-space. The 2D projection after 20 iterations of adversarial generation followed by 4-class LDA is shown in Figure 3.6C, with the magenta decoy data clustering around the center of the 2D LDA-space. Adversarially generated decoy spectra were accumulated in each iteration. Results from 80 iterative loops are shown in Figure 3.6D with the magenta corresponding to early iterations and yellow to subsequent iterations. The separation between genuine and decoy data along the LD1 coordinate is most evident in the training data.

3.3.3 Resolution Convergence

Using resolution defined earlier in Equation 3.1, an inspection of how this metric evolves during the iterative process is plotted in Figure 3.7A representative comparison of the multi-class resolution J_{tot} achieved by PCA and GALDA is shown in Figure 3.7A. The trend in the GALDA curves demonstrates a smooth convergence between the testing and the training

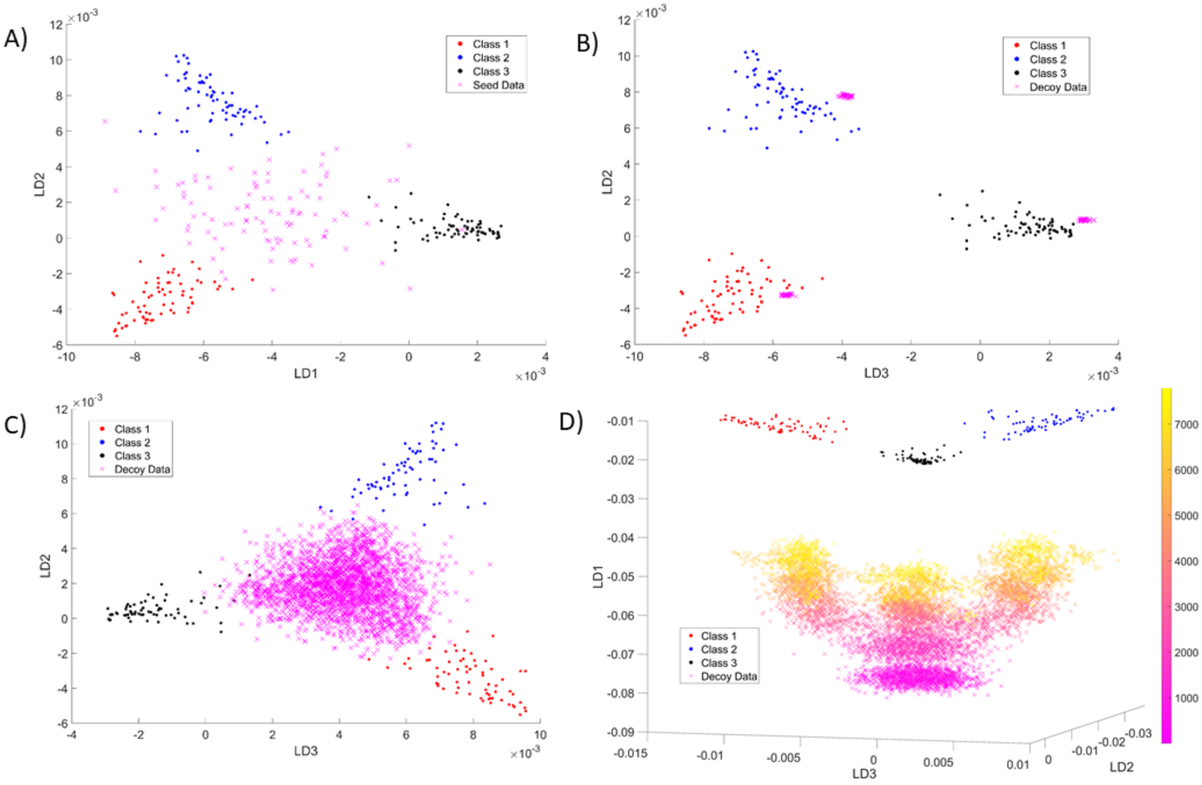


Figure 3.6. An overview of the GALDA process. (A) 3 class LDA results together with random initial seed spectra (magenta). (B) Seed spectra following perturbation to produce decoy data (magenta) projected using the same reduction vectors in (A), C) 4-class LDA with the same data in (B), with the decoy data (magenta) separated along a third LDA coordinate into the page. (D) 3D rendering of a 4-class LDA (3 genuine + 1 decoy) after 80 iterations, illustrating the separation of genuine and decoy spectra by GALDA.

data. In the early iterations of GALDA, the large gap in the multi-class resolution between the training and testing data is indicative of strong overfitting, in which both signal and spurious noise contribute to the resolution within the training data. For comparison, PCA reduction to 2D PCA-space using the training data is displayed as dashed lines for both training and testing data. This trend is consistent when performing five-fold cross validation.

This study with just three training spectra as shown in Figure 3.7B suggests that mathematical instabilities preventing the use of LDA at full dimension no longer serve as a barrier for analysis via GALDA. In practice, the use of three training spectra is highly inadvisable, as such a small sampling size is generally unlikely to represent a sufficient population of spectra for reliable classification. Nevertheless, the successful mathematical implementation of GALDA in this extreme limit supports the assertion that non-representative sampling, rather than instabilities in the algorithm, serves as the practical bottleneck dictating the reliability of GALDA in the limit of small training size.

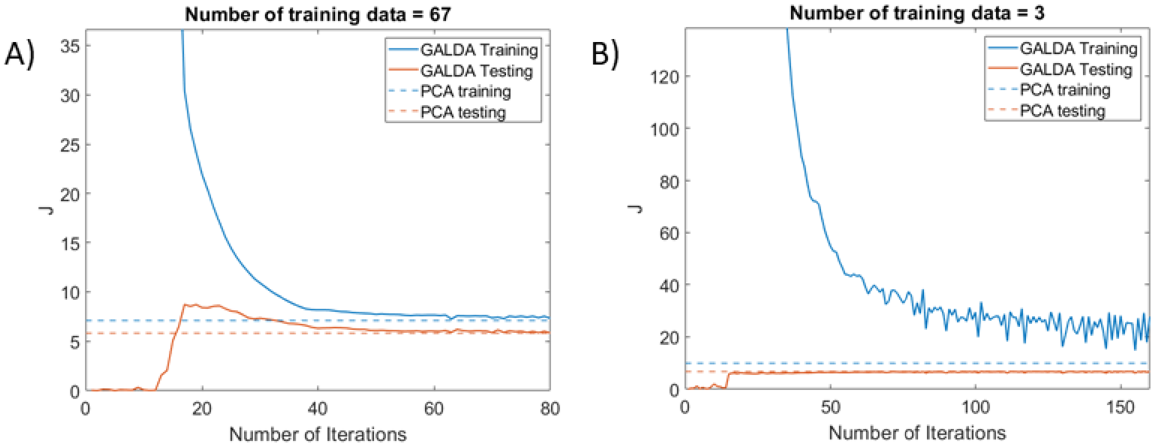


Figure 3.7. Representative trends in resolution J_{tot} evaluated with increasing iterations of GALDA, with PCA results shown as comparators; (A) 80% training, 20% testing (B) 3% training, 97% testing.

3.3.4 Cross-Validated Comparison With Other Methods

The performance of GALDA was compared with other common dimension reduction methods of spectral data, the results of which are summarized in Figure 3.8. Specifically, alternative methods included PCA, PLS-DA, PCA+LDA, and regularized LDA using a shrunken centroids regularizer[44]. These methods were selected based on their simplicity and ubiquity and are not meant to represent an exhaustive set of comparators for dimension reduction. With the exception of PCA, the others are also supervised methods that retain the label information when performing dimension reduction, analogous to GALDA. Two key metrics were considered for assessing the merits of the different dimension reduction approaches: resolution (or more properly, the squared resolution given by J) and overfitting. Resolution in terms of J_{tot} is defined in detail in Eq. 3.1. Overfitting is simply given by the difference in J_{tot} between the testing and training data. Five-fold cross-validation was performed, with the means and standard deviations given in Figure 3.8. Two datasets with different sample distribution were used to evaluate the different models mentioned above. The first row correspond to results from Raman simulations derived from a database of mineral spectra and designed with bimodal distributions within each class. The second row correspond to experimentally measured Raman data of clopidogrel bisulfate.

The differences in testing set resolutions indicated in Figure 3.8 reveal interesting, but not entirely unexpected trends. For the clopidogrel bisulfate data, all methods considered yielded only minor differences in the recovered resolution following five-fold cross-validation. For the simulated dataset with a non-monotonic probability distribution, the resolution for PCA was well below that of all the supervised methods. This result is entirely consistent with differences expected for supervised versus unsupervised dimension reduction; unsupervised methods neglect to incorporate class information in the dimension reduction process. As such, PCA and other unsupervised methods are well-established to be prone to underperform using supervised datasets. A representation of the simulated spectra in a 2D reduced dimensional PCA-space was shown in Figure 3.4, providing a clear visual indication of the mechanism for resolution failure when maximizing the total variance in the data in PCA irrespective of class status. Noteworthy in the simulated dataset, PCA+LDA and RLDA

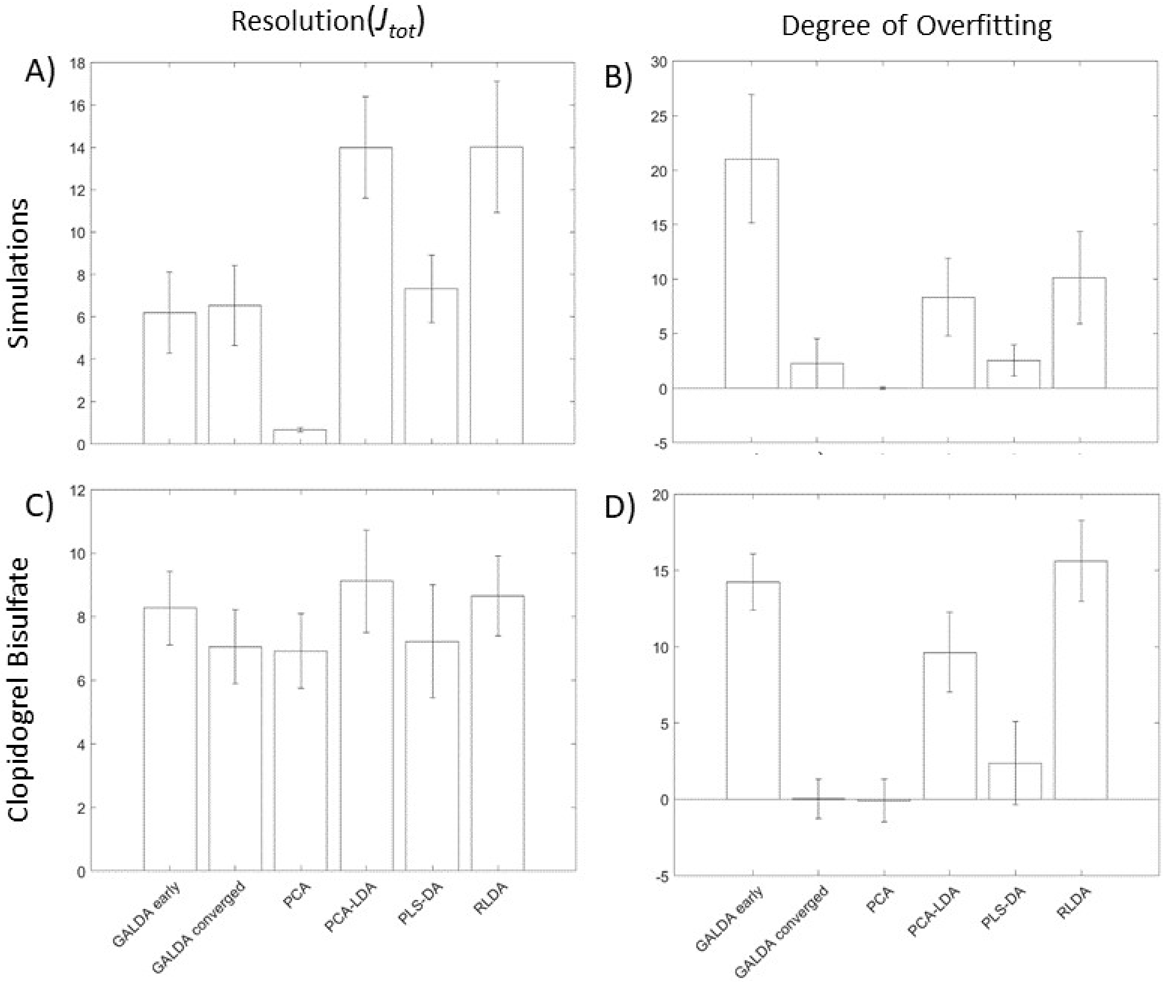


Figure 3.8. Cross-validated comparison results of GALDA with other standard methods, tested with two datasets. The first row is for a simulated dataset, while the second row is calculated for experimental Raman spectra of clopidogrel micro spheroids. Metrics are resolution (first column) and the degree of overfitting (second column). Unit-less resolution is defined by Fisher's discriminant J_{tot} using Eq. 3.1. The degree of overfitting is given by the difference in J_{tot} between the training and testing data.

yielded resolutions for the testing data that were statistically higher than GALDA and PLS-DA. Initially, one might be be drawn by this outcome to conclude that PCA+LDA and RLDA should be preferred as dimension reduction methods relative to GALDA. However, resolution alone is a potentially hazardous metric without also considering overfitting.

The overfitting trends shown in Figure 3.8 are particularly noteworthy. First and foremost, GALDA and PCA consistently produced the smallest degrees of overfitting relative to all other methods considered for both simulated and genuine spectra, with PCA clearly outperforming GALDA. Second, the gains in testing resolution from PCA+LDA and RLDA discussed in the preceding paragraph are generally offset by corresponding increases in overfitting. As discussed in the Introduction, high overfitting susceptibility increases the likelihood of low resolution and/or misclassification for data outside the testing and training datasets. Both PCA+LDA and RLDA yielded overfitting substantially greater than the resolution for the clopidogrel bisulfate dataset and on par with the resolution for the simulations. Consistent with this propensity, PCA+LDA and RLDA yielded high roughness coefficients in the corresponding loading plots, with the potential to integrate residual noise contributions in the dimension reduction operation. In the converse extreme, the reduced overfitting by PCA relative to GALDA is also offset by the potential for major reductions in resolution, as illustrated in the simulated dataset. As such, GALDA and PLS-DA strike an attractive balance between maximizing resolution while minimizing overfitting, with GALDA providing comparable resolution and lower overfitting than PLS-DA in the test cases considered herein.

The improved performance of GALDA relative to PCA+LDA is particularly noteworthy, since both algorithms conclude with an LDA operation. For GALDA, LDA is performed at full dimension of p including both genuine and generated data, while for PCA+LDA, LDA is performed at the full dimension n of PCA, which is lower than p. The improved performance of GALDA may arise from the retention of information that is is necessarily lost in the initial dimension reduction to PCA-space in PCA+LDA. Alternatively, improvements by GALDA may arise from iterative suppression of residual overfitting still present within the PCA+LDA operation and not removed by the initial transformation to PCA-space. Since

PCA+LDA and GALDA yielded similar resolutions but substantial differences in overfitting, we attribute the improvement to the latter of the two explanations.

3.3.5 Mechanism Discussion

Hints on the mechanism of action for achieving this suppression of overfitting by GALDA can be gleaned from inspection of the 3D projection shown in Figure 3.6D, which includes both genuine and decoy data. Interestingly, decoy data in the initial iterations were centrally located consistent with an entirely random distribution, with subsequent iterations gravitating towards the central regions of each of the three classes in all three LDA dimensions. This trend is apparent from the gradient color of the decoy dataset, changing from magenta to yellow, with the color bar representing the number of seed/decoy data accumulated. Consideration of the collective data set resulted in an inverted "three-legged stool" shape for the generated decoy data. Qualitatively, we interpret this trend to indicate that initial iterations resulted in general suppression of noise within the loading plots, subsequently transitioning to the production of generated decoy data increasingly similar in form to the genuine data.

In this implementation of GALDA, the first LDA axis (LD1) always spontaneously corresponded to the direction of greatest separation for distinguishing genuine vs. generated spectra. It is worthwhile considering the generality of this assignment; it is conceivable that an axis within the genuine data set may potentially exhibit a greater eigenvalue for the Fisher linear discriminant than that arising from the disparity between genuine and decoy data, leading to mis-assignment of the generative adversarial discriminatory axis. Fortunately, several safeguards can be built into the GALDA algorithm to ensure that LD1 is universally selected as the generative adversarial discriminatory coordinate. In the simplest, the weighting of the adversarial data can be rescaled within the LDA assessment to place a greater discriminatory weight on the generated decoy data. This rescaling can be performed either by multiplication of the class variance (both within and between) of the generated decoy data by a scalar greater than 1, or by removal of normalizations for the number of spectra in each class. In the latter case, the larger accumulated number of generated data alone are sufficient to ensure assignment to LD1.

3.3.6 Visualization of the Loading Vectors

A comparison of the loading vectors for dimension reduction to 2D by PCA and GALDA for the clopidogrel bisulfate data is shown in Figure 3.9, indicating close agreement between the two methods. In general, one would not expect GALDA and PCA to produce similar loading plots. The PCA axes optimally maximize the total variance within the data set, while LDA axes are generated to optimize the resolution between the classes. The slight but statistically significant improvement in resolution from GALDA relative to PCA within the testing data is likely a consequence of the different objective functions between LDA and PCA, as LDA explicitly optimizes resolution. The convergence of GALDA to a result close to PCA is reassuring, as PCA is anticipated to perform reliably in the limits of small class numbers and small training set size of monotonically distributed data. Also consistent with expectations, the loading vectors for full dimension LDA prior to iterative optimization were dominated by noise, suggestive of high over-fitting. Inspection of the mean spectra for each class in Figure 3.9B provides context for interpreting the features recovered by GALDA. For example, in Figure 3.9A loading vector 2, both PCA and GALDA recover two peaks shouldering each other at 380 cm^{-1} and a single peak at 1230 cm^{-1}, with a sharp peak at 380 cm-1 in loading vector 3. Not surprisingly, these same peaks differ notably in the mean spectra for the three different classes, displayed in Figure 3.9B.

3.3.7 Loading Vectors Roughness Evaluation

If it is reasonably assumed that the key spectral features supporting discrimination between the different Raman data sets generally consist of peaks spanning multiple pixels, then one would expect optimal loading plots to reflect this trend, with minimal high-frequency content. While not as robust as metrics based on resolution and overfitting, smoothness has the distinct advantage of not being a property explicitly optimized by any of the algorithms considered. As such, it offers the prospect of a low-bias assessment of the quality in the recovered loading plots. To quantify this property such that the smoothness of the loading vector produced by different models can be compared against one another, the contrary of smoothness – roughness – is computed, as it is mathematically easier to define. Given a

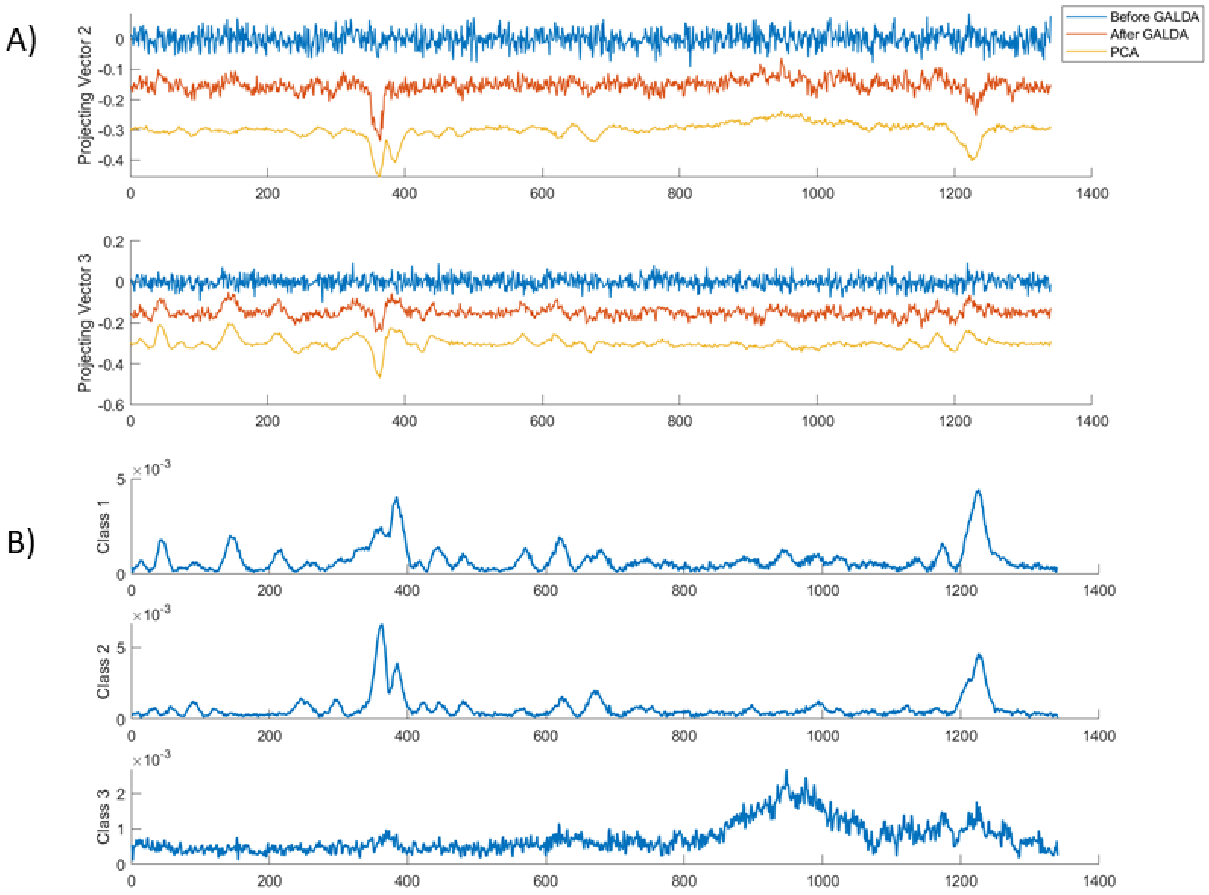


Figure 3.9. (A) Before and after GALDA loading vectors comparison. The PC loadings are plotted as well for reference. (B) Mean spectrum of the original data.

vector $\boldsymbol{v}$, roughness can be defined as the root-sum-of-squares (RSS) of the second derivative of $\boldsymbol{v}$, evaluated at the highest accessible frequency (i.e., by convolution with a 3-element derivative digital filter) and normalized by the RSS of $\boldsymbol{v}$:

$$\text{roughness} = \frac{\left\|\overrightarrow{\boldsymbol{v}\prime\prime}\right\|_2}{\left\|\overrightarrow{\boldsymbol{v}}\right\|_2} \tag{3.2}$$

Quantitative evaluation of the roughness is shown in Figure 3.10 for the suite of dimension reduction methods considered herein. The roughness generally follows the overfitting trends summarized in Figure 3.8, consistent with residual high-frequency noise in the loading plots coinciding with a higher overfitting propensity.

3.3.8 Optimization: PLS-DA and Regularization

As a complement to the core focus on the simplest and most common dimension reduction methods (LDA and PCA), studies were also performed to assess GALDA relative to dimension reduction by partial least squares discriminant analysis (PLS-DA). PLS-DA has greater computational stability than LDA, and unlike PCA, retains classification information for supervised learning. In addition, PLS-DA is readily available within many widely available data analysis packages. In the present study, PLS-DA was performed using an open source MATLAB GUI tool described by Zontov *et al.*[45]. Two separate evaluations for PLS-DA were performed and are shown in Figure 3.12. In the first plot (Figure 3.12A), PLS-DA was performed with the recommended default 12 PLS components for a 12-dimensional PLS-DA space, the first two of which producing the greatest resolution in the training data are displayed in the figure. As a complement, PLS-DA was also performed with just two PLS components to match the 2D dimensionality in which the data are displayed in Figure 3.12B.

The trends depicted share qualitative similarities to those in Figure 3.5. When evaluated with many components, PLS-DA produces large disparities in the resolution produced in the training data relative to the testing data, consistent with the over-fitting observed in conventional LDA. In contrast, the PLS-DA performed with just two components resulted

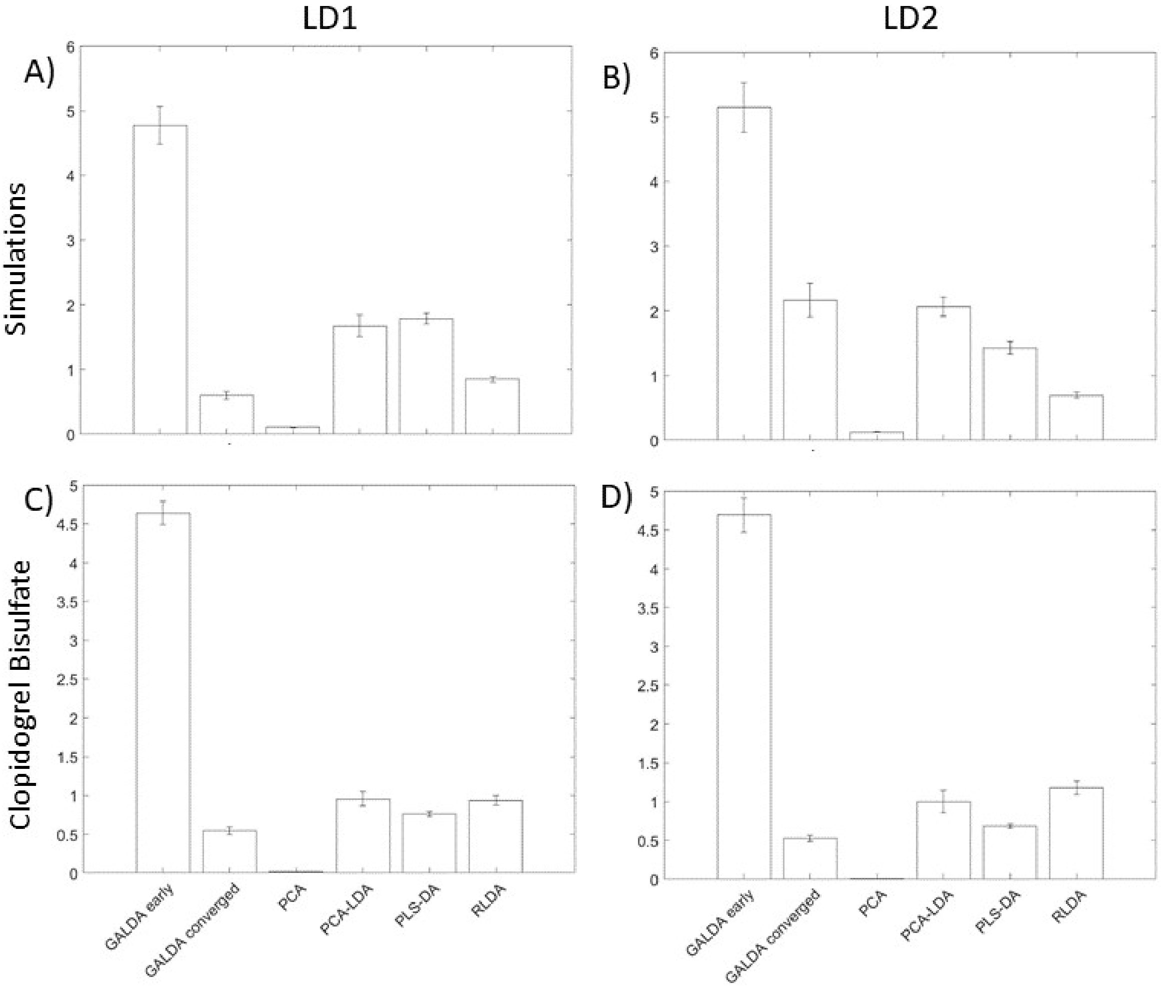


Figure 3.10. Evaluation of the roughness of the loading vectors with different dimension reduction algorithms.

in good agreement between the testing and training data, but with a total resolution J_{tot} lower than both PCA and GALDA under identical conditions (former resulted in a J_{tot} of around 6.0 while the latter around 6.8).

Both of these trends can be easily rationalized by consideration of the limiting behaviors of PLS-DA evaluated in the extremes of high and low dimension as detailed nicely in a 2014 tutorial review by Brereton and Lloyd[19]. In the extreme of full dimension (in this case, a total of 1340 PLS-DA components), PLS-DA becomes mathematically identical to the full-dimension LDA employed in the present study. While 12 is certainly much less than 1340, the general principle still applies. When evaluated with a sufficiently large number of components, PLS-DA becomes prone to the same over-fitting instabilities inherent within LDA. In this limit, it is therefore not surprising to observe large disparities between the resolutions achievable in the testing and training data and correspondingly low resolution within the training data set. Indeed, these trends strongly suggest that generative adversarial approaches may greatly improve the general reliability of PLS-DA through reduction of over-fitting contributions by analogy with GALDA. The lower resolution of PLS-DA relative to LDA and PCA when evaluated with two components can also be rationalized by the extreme limits of a single PLS-DA component. As demonstrated by Brereton and Lloyd[19], PLS-DA converges to a projection of the data along a vector paralleling the simple Euclidean distance to centroids (EDC) in a two-class system. Dimension reduction based on the EDC incorporates no information on the variance about the centroid, which plays a major role in optimizing the resolution. As such, the total resolution is anticipated to be less than that recovered by LDA.

Regularization based on the shrunken centroids regularized discriminant analysis (RDA) was implemented by scaled addition of an identity matrix to $\boldsymbol{S_W}$, and the scaling was adjusted within the range between 0 and 1, following Guo, Hastie, and Tibshirani[16]. Depending on the scaling constant, the rate of the convergence was affected, as illustrated in Figure 3.11. From the trends shown, regularization clearly has a major effect in the early iterations of GALDA but exerts little influence on the final outcome following convergence, after which LDA is no longer ill-posed. It should be noted that the shrunken centroids approach to RDA is just one relatively simple example of a larger class of regularization architectures.

The integration of RLDA with GALDA resulted in significant improvements in the rate of convergence in GALDA with similar outcomes in resolution and overfitting. The reduction in overfitting afforded by GALDA to enable full-dimension LDA with $p > n$ suggests that generative adversarial data augmentation strategies may help suppress overfitting in broader classes of classified dimension reduction and discrimination algorithms.

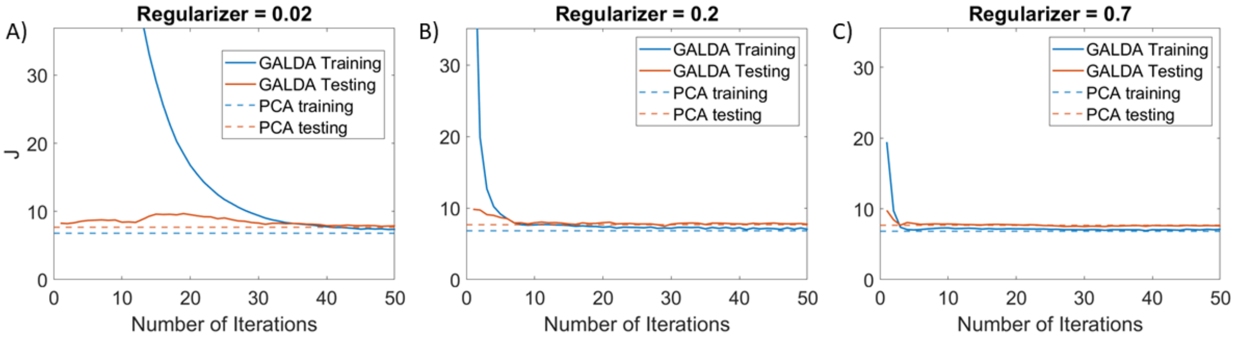


Figure 3.11. Regularized LDA with different scaling constants, evaluated for the clopidogrel bisulfate Raman data split between 80% training and 20% testing.

3.3.9 Significance of GALDA

Given the conceptual origin of GALDA emerging from the success of GANs in artificial neural networks, it is interesting to compare and contrast the two data-mining tools. The established utility of GANs for data augmentation to minimize over-fitting in ANNs is directly analogous to the reductions in over-fitting realized in GALDA relative to naive LDA[46]–[48]. In addition, back-propagation algorithms for optimization of attacks in GANs have direct analogies to the analytical expression derived in this work for the optimized perturbations in GALDA, both of which are based on gradient approaches. However, despite these high-level similarities, the implementation and output of GANs and GALDA are notably distinct. Perhaps most strikingly, GALDA has the distinct advantage of linearity in design; the final product of GALDA is a relatively simple linear transformation of the data

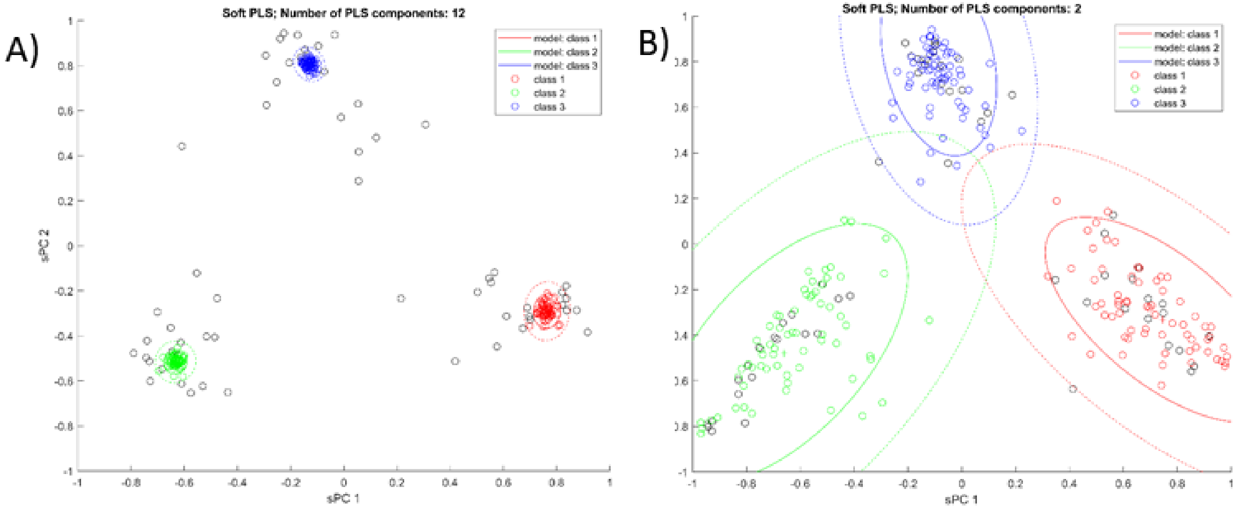


Figure 3.12. Dimension reduction by PLS-DA for the same Raman data set for reference. Selecting a default number of 12 PLS components with 65 training spectra as shown in A) leads to large differences in resolution between the testing (RGB circles) and training (black circles) data. Reducing the components to two recovers trends qualitatively similar to those produced by PCA.

to produce continuous distributions of transformed data. Uncertainties can be propagated straightforwardly through linear transformations associated with dimension reduction by the methods considered herein, with common statistical methods for analysis and testing still firmly holding. The assumption of linearity serves as an additional constraint in GALDA and the other dimension reduction methods considered, enabling reliable implementation with much smaller training set sizes relative to analogous neural network architectures. Of course, these strengths only hold when data are linearly connected; nonlinear relationships (e.g., image magnification, rotation, translation, etc.) are difficult to capture by any of the methods considered, in which case transitioning to ANN architectures may improve classification accuracy. Additionally, most GAN-assisted ANN platforms produce discretized decisions as outputs (e.g., class determination of a given input) rather than used for dimension reduction/data visualization. As such, the machine learning engine driving GALDA can be viewed as highly complementary to GAN platforms, implementing generative adversarial strategies in ways inaccessible to existing neural network architectures.

3.4 Conclusions

This chapter described an iterative generative adversarial approach to address overfitting propensities inherent in linear discriminant analysis at full dimension and assess key figures of merit in comparison with PCA, PLS-DA, PCA+LDA, and RDA methods using both simulated and experimentally measured Raman spectral data sets. Relative to other supervised dimension reduction approaches considered, GALDA yielded substantial reductions in overfitting and comparable resolution. Overfitting using GALDA was generally comparable to or higher than PCA, but with substantial gains in resolution by GALDA in simulations with multi-polar distributions, consistent with expectations for unsupervised classifiers. These results suggest potential broad utility of GALDA to minimize overfitting in full-dimension LDA of supervised spectral datasets. The substantial reductions in overfitting arising in GALDA relative to other supervised methods considered herein for dimension reduction methods are tentatively attributed to the targeted suppression of only the subset of noise features that overlap with genuine spectra in the reduced dimensional space. Despite the successes of GALDA in this preliminary study, readers should be cautioned that the comparators and GALDA were all performed using default parameters to minimize introduction of complicating meta-parameters; as such, further optimization of the comparator methods could reduce potential benefits by GALDA suggested by the current study.

References

[1] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, *The rise of "big data" on cloud computing: Review and open research issues*, Jan. 2015. DOI: 10.1016/j.is.2014.07.006.

[2] C. L. Philip Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences*, vol. 275, pp. 314–347, Aug. 2014, ISSN: 00200255. DOI: 10.1016/j.ins.2014.01.015.

[3] I. K. Fodor, "A survey of dimension reduction techniques," *Library*, vol. 18, no. 1, pp. 1–18, May 2002. DOI: 10.2172/15002155. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.122.201%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf.

[4] M. Carreira-Perpinán, "A review of dimension reduction techniques," *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, pp. 1–69, 1997. [Online]. Available: http://www.pca.narod.ru/DimensionReductionBrifReview.pdf.

[5] M. Bahri, A. Bifet, S. Maniu, and H. M. Gomes, "Survey on Feature Transformation Techniques for Data Streams," Tech. Rep., 2020, pp. 4796–4802. DOI: 10.24963/ijcai.2020/668.

[6] G. Chao, Y. Luo, and W. Ding, "Recent Advances in Supervised Dimension Reduction: A Survey," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 341–358, Jan. 2019, ISSN: 2504-4990. DOI: 10.3390/make1010020. [Online]. Available: https://www.mdpi.com/2504-4990/1/1/20.

[7] P. Benner, D. C. Sorensen, and V. Mehrmann, Eds., *Dimension Reduction of Large-Scale Systems*, ser. Lecture Notes in Computational Science and Engineering. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, vol. 45, ISBN: 978-3-540-24545-2. DOI: 10.1007/3-540-27909-1. [Online]. Available: http://link.springer.com/10.1007/3-540-27909-1.

[8] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea, "Towards a Quantitative Survey of Dimension Reduction Techniques," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, Sep. 2019, ISSN: 1077-2626. DOI: 10.1109/tvcg.2019.2944182.

[9] S. Balakrishnama and A. Ganapathiraju, "Linear Discriminant Analysis - a Brief Tutorial," Ph.D. dissertation, 1998.

[10] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, Feb. 2004, ISSN: 0047259X. DOI: 10.1016/S0047-259X(03)00096-4.

[11] A. Sharma and K. K. Paliwal, "Linear discriminant analysis for the small sample size problem: an overview," *International Journal of Machine Learning and Cybernetics*, vol. 6, no. 3, pp. 443–454, 2015. DOI: 10.1007/s13042-013-0226-9.

[12] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small sample size problem of LDA," in *Proceedings - International Conference on Pattern Recognition*, vol. 16, 2002, pp. 29–32. DOI: 10.1109/icpr.2002.1047787.

[13] L. F. Chen, H. Y. M. Liao, M. T. Ko, J. C. Lin, and G. J. Yu, "New LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000, ISSN: 00313203. DOI: 10.1016/S0031-3203(99)00139-9.

[14] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI Communications*, vol. 30, no. 2, pp. 169–190, 2017, ISSN: 09217126. DOI: 10.3233/AIC-170729.

[15] S. J. Dixon and R. G. Brereton, "Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on," *Chemometrics and Intelligent Laboratory Systems*, vol. 95, no. 1, pp. 1–17, 2009, ISSN: 01697439. DOI: 10.1016/j.chemolab.2008.07.010. [Online]. Available: http://dx.doi.org/10.1016/j.chemolab.2008.07.010.

[16] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, no. 1, pp. 86–100, Jan. 2007, ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxj035. [Online]. Available: https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxj035.

[17] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989, ISSN: 1537274X. DOI: 10.1080/01621459.1989.10478752. [Online]. Available: https://www.tandfonline.com/action/journalInformation?journalCode=uasa20.

[18] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 3, pp. 862–873, Mar. 2009, ISSN: 01962892. DOI: 10.1109/TGRS.2008.2005729.

[19] R. G. Brereton and G. R. Lloyd, "Partial least squares discriminant analysis: Taking the magic away," *Journal of Chemometrics*, vol. 28, no. 4, pp. 213–225, Apr. 2014, ISSN: 1099128X. DOI: 10.1002/cem.2609. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/cem.2609%20https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.2609%20https://onlinelibrary.wiley.com/doi/10.1002/cem.2609.

[20] L. C. Lee, C. Y. Liong, and A. A. Jemain, "Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps," *Analyst*, vol. 143, no. 15, pp. 3526–3539, Aug. 2018, ISSN: 13645528. DOI: 10.1039/c8an00599k. [Online]. Available: https://pubs.rsc.org/en/content/articlehtml/2018/an/c8an00599k%20https://pubs.rsc.org/en/content/articlelanding/2018/an/c8an00599k.

[21] Y. Mallet, D. Coomans, and O. De Vel, "Recent developments in discriminant analysis on high dimensional spectral data," *Chemometrics and Intelligent Laboratory Systems*, vol. 35, no. 2, pp. 157–173, Dec. 1996, ISSN: 01697439. DOI: 10.1016/S0169-7439(96)00050-0.

[22] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, no. 1, pp. 86–100, 2007.

[23] D. M. Witten and R. Tibshirani, "Penalized classification using fisher's linear discriminant," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 5, pp. 753–772, 2011.

[24] P. J. Bickel, E. Levina, *et al.*, "Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, no. 6, pp. 989–1010, 2004.

[25] J. J. Workman, *A Review of Calibration Transfer Practices and Instrument Differences in Spectroscopy*, Mar. 2018. DOI: 10.1177/0003702817736064. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0003702817736064.

[26] R. N. Feudale, N. A. Woody, H. Tan, A. J. Myles, S. D. Brown, and J. Ferré, *Transfer of multivariate calibration models: A review*, Nov. 2002. DOI: 10.1016/S0169-7439(02)00085-0.

[27] S. F. C. Soares, A. A. Gomes, M. C. U. Araujo, A. R. G. Filho, and R. K. H. Galvão, *The successive projections algorithm*, Jan. 2013. DOI: 10.1016/j.trac.2012.09.006.

[28] L. Li, X. Pan, W. Chen, M. Wei, Y. Feng, L. Yin, C. Hu, and H. Yang, "Multi-manufacturer drug identification based on near infrared spectroscopy and deep transfer learning," *Journal of Innovative Optical Health Sciences*, vol. 13, no. 04, p. 2 050 016, Jul. 2020, ISSN: 1793-5458. DOI: 10.1142/S1793545820500169. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S1793545820500169.

[29] R. Zhang, H. Xie, S. Cai, Y. Hu, G.-k. Liu, W. Hong, and Z.-q. Tian, "Transfer-learning-based Raman spectra identification," *Journal of Raman Spectroscopy*, vol. 51, no. 1, pp. 176–186, Jan. 2020, ISSN: 0377-0486. DOI: 10.1002/jrs.5750. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/jrs.5750.

[30] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative Adversarial Networks for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018, ISSN: 01962892. DOI: 10.1109/TGRS.2018.2805286.

[31] S. Yu, H. Li, X. Li, Y. V. Fu, and F. Liu, "Classification of pathogens by Raman spectroscopy combined with generative adversarial networks," *Science of the Total Environment*, vol. 726, p. 138 477, Jul. 2020, ISSN: 18791026. DOI: 10.1016/j.scitotenv.2020.138477.

[32] I. V. Tetko, D. J. Livingstone, and A. I. Luik, "Neural Network Studies. 1. Comparison of Overfitting and Overtraining," *Journal of Chemical Information and Computer Sciences*, vol. 35, no. 5, pp. 826–833, 1995, ISSN: 00952338. DOI: 10.1021/ci00027a006. [Online]. Available: https://pubs.acs.org/sharingguidelines.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Tech. Rep., 2012, pp. 1097–1105. [Online]. Available: http://code.google.com/p/cuda-convnet/.

[34] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, Nov. 2016. arXiv: 1611.03530. [Online]. Available: http://arxiv.org/abs/1611.03530.

[35] M. Claesen and B. De Moor, "Hyperparameter search in machine learning," *arXiv preprint arXiv:1502.02127*, Feb. 2015. arXiv: 1502.02127. [Online]. Available: http://arxiv.org/abs/1502.02127.

[36] W. Zhao, "Research on the deep learning of the small sample data based on transfer learning," in *AIP Conference Proceedings*, vol. 1864, American Institute of Physics Inc., Jul. 2017, p. 020 018, ISBN: 9780735415423. DOI: 10.1063/1.4992835. [Online]. Available: http://aip.scitation.org/doi/abs/10.1063/1.4992835.

[37] I. Goodfellow, P.-A. Jean, M. Mehdi, X. Bing, W.-F. David, O. Sherjil, C. Aaron, B. Yoshua, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 3, no. January, pp. 2672–2680, 2014, ISSN: 10495258. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[38] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" In *Advances in Neural Information Processing Systems*, January, vol. 4, 2014, pp. 3320–3328. arXiv: 1411.1792.

[39] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations," Tech. Rep., Feb. 2019, pp. 686–707. DOI: 10.1016/j.jcp.2018.10.045. arXiv: 1711.10561v1.

[40] I. Goodfellow, "Tutorial: Generative Adversarial Networks," in *Proceedings of Neural Information Processing Systems*, Dec. 2016. arXiv: 1701.00160. [Online]. Available: http://arxiv.org/abs/1701.00160%20http://arxiv.org/abs/1701.00160..

[41] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications," vol. 14, no. 8, Jan. 2020. arXiv: 2001.06937. [Online]. Available: http://arxiv.org/abs/2001.06937.

[42] A. B. Nicolae BUZGAR, Andrei Ionut APOPEI, *Romanian Database of Raman Spectroscopy*, 2009. [Online]. Available: http://rdrs.ro.

[43] R. Liu and D. F. Gillies, "Overfitting in linear feature extraction for classification of high-dimensional image data," *Pattern Recognition*, vol. 53, pp. 73–86, 2016, ISSN: 00313203. DOI: 10.1016/j.patcog.2015.11.015.

[44] A. Biancolillo and F. Marini, *Chemometric methods for spectroscopy-based pharmaceutical analysis*, 2018. DOI: 10.3389/fchem.2018.00576.

[45] Y. V. Zontov, O. Y. Rodionova, S. V. Kucheryavskiy, and A. L. Pomerantsev, "PLS-DA – A MATLAB GUI tool for hard and soft approaches to partial least squares discriminant analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 203, p. 104 064, Aug. 2020, ISSN: 18733239. DOI: 10.1016/j.chemolab.2020.104064.

[46] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101 552, Dec. 2019, ISSN: 13618423. DOI: 10.1016/j.media.2019.101552. arXiv: 1809.07294.

[47] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *International workshop on simulation and synthesis in medical imaging*, Springer, 2018, pp. 1–11.

[48] A. Gooya, O. Goksel, I. Oguz, and N. Burgos, *Simulation and Synthesis in Medical Imaging*, 2018. DOI: 10.1007/978-3-030-00536-8.

4. PHYSICS ENCODED NEURAL NETWORK (PENN) FOR FLUORESCENCE RECOVERY AFTER PHOTOBLEACHING DATA ANALYSIS

4.1 Introduction

Diffusion measurements of protein in solutions or matrices bear great importance concerning drug release/deliver [1], even rational drug design[2]. Given the importance and urgency, multi-dimensional research has been conducted for measuring, modeling and probing the diffusion process[3], [4]. On the experimental side, new methodologies are being proposed to aid the measurements of diffusion coefficient, which in turn contribute to screening drugs and modified drug conjugates[5], [6]. On the theory side, models have been created to simulate the process of drug-releasing[7]. With the advances of artificial intelligence and machine learning methods, it's attempting to explore whether these methods apply to the area of protein diffusion measurement, as artificial neural networks (ANNs) have a proven record of predicting outcomes with high precision albeit being a black box[8]–[10].

For the success of ANNs to translate into protein interactions, it would be negligent to treat data as they are in digital image processing or natural language processing, which is the primary area ML methods demonstrated success. The architecture is hardly a naive translation, i.e. treating feature space as randomly pointing vectors. Protein interactions entail mathematical equations regulated by the fundamental law of forces that connect features. However, a "black-box" neural networks obfuscate the underlying physics driving the outcome prediction. Furthermore, the data requirements for implementation substantially increase the expense and complexity in acquiring protein-related training data. This chapter proposes the integration of physics-based constraints with tunable generalizability directly into the fabric of an ANN as an intermediate layer, thereby reducing data volumes needed for training and providing access to physical insights into the driving mechanisms in the diffusion model for proteins.

Of the research involving machine learning and protein interaction, the focus has been primarily on optimizing the set of hyper-parameters such that existing ML methods can produce high-quality results[11]. The integration of physics-based parameters into ML is less extensive and quite new in comparison. The basic idea is to incorporate physical laws to restrict/inform neural networks with the need for less data for reliable model training. Raissi, Perdikaris, and Karniadakis[12] first introduced the concept of using neural networks as universal function approximators, as an alternative to substitute the classical numerical method for solving partial differential equations (PDEs). Physics informed neural networks (PINNs) are the then employed to fluid dynamics[13]–[15] and blood pressure prediction[16], which were usually achieved by driving the neural network to minimize the residual of physics questions governing the physical process; Baldi *et al.*[17] proposed a new strategy named parameterized neural network in which the approach is slightly different, the list of input features were extended to include more parameters for describing connected physics.

The work in this chapter is an extension to that of Baldi *et al.* A physics parameter that is the dependent variable of the rest of the inputs variable is included in the input features; the output of the ANN is a set of physical parameters that can be used to mathematically calculate the dependent variable. The structure is validated first using simulated diffusion data, then experimental FRAP data. The motivation for the design of a custom ANN is two-fold: i) it can be purpose-built for our specific physics-based models for intermolecular interactions, and ii) the additional constraints imposed by physics will reduce the size of the required training data.

4.2 Methods

4.2.1 DLVO Theory

Starting from a DLVO framework to extend the perturbations to diffusion coefficients from protein-protein self-interactions to protein-matrix interactions[18], [19]:

$$D = D_0 \left(1 + k_{D_{22}} c_2 + k_{D_{23}} c_3\right) \tag{4.1}$$

In this model, the protein/matrix interaction based on their concentrations (c_2, c_3) deviates the diffusion coefficient D from the asymptotic diffusion coefficient D_0, characterized by respective factors ($k_{D_{22}}$ for self interactions, $k_{D_{23}} for protein - matrix interactions$). For the present purposes, the concentration of matrix is anticipated to greatly exceed that of the protein, such that the corrections from protein-matrix interactions are anticipated to exceed those of the corrections from variation in protein concentration. This assumption is further justified by the absence of a protein concentration gradient in FRAP measurements, in which the composition of the droplets is generally spatially uniform prior to analysis. In all instances presented herein, the protein concentration was identical in each experiment to isolate changes due just to perturbations from $k_{D_{23}}$. In the absence of intermolecular interactions, the asymptotic diffusion coefficient D_0 can be described using the Stokes-Einstein equation based on the hydrodynamic radius of the protein R_h^a and the viscosity η.

$$D_0 = \frac{k_B T}{6\pi\eta R_h^a} \tag{4.2}$$

In Eq. 4.2, k_B is the Boltzmann constant and T is the temperature.

The matrix concentration-dependent perturbation to the diffusion coefficient $k_{D_{23}}$ can be written in terms of two main contributions from excluded volume $k_{D_{23}}^{\mathrm{ex}}$ and short-range interactions $k_{D_{23}}^{sr}$.

$$k_{D23} = k_{D_{23}}^{\mathrm{ex}} + k_{D_{23}}^{sr} \tag{4.3}$$

The excluded volume contribution $k_{D_{23}}^{\mathrm{ex}}$ is directly proportional to the corresponding second virial cross-coefficient contribution from excluded volume b_{23}^{ex}. Consistent with the notation of Roberts *et al.*[20], the virial coefficient expressed in units of volume (inverse number density) will be indicated with lowercase notation. Conversion to the virial coefficient in terms of concentration normalized to molecular weight in units of $\mathrm{mL\,g^{-2}}$ indicated by B_{23}^{ex} can be done by the following relation connecting the two different common units conventions.

$$B_{23} = \frac{b_{23} N_A}{M^a M^b} \tag{4.4}$$

In the preceding expression, N_A is Avogadro's number, M is the molecular weight of either the protein (superscript a) or the polymer (superscript b). The expression for b_{23}^{ex}

can be calculated analytically based on a hard-sphere model, where R_h is the hydrodynamic radius for either the protein (superscript a) or the polymer (superscript b).

$$b_{23}^{\mathrm{ex}} = \frac{4\pi\left(R_h^a + R_h^b\right)^3}{3} \tag{4.5}$$

The expression for the excluded volume contribution to the diffusion constant b_{23}^{ex} is derived from the sum of virial coefficients that are evaluated over the integration bounds of $r = 0$ to $r = 2R_h$[20], [21].

$$k_{D23}^{\mathrm{ex}} = 0.39 B_{23}^{\mathrm{ex}} M^b \tag{4.6}$$

The expression for the short-range contribution to the diffusion constant $k_{D_{23}}^{sr}$ is derived from the sum of virial coefficients that are evaluated over the integration bounds of $r = 2R_h$ to $r = \infty$.

$$k_{D23}^{sr} = -1.024(B_{23}^{\mathrm{ex}} - B_{23})M^b \tag{4.7}$$

The virial coefficient, B_{23}, can be defined as the sum of excluded volume, electrostatic, and short-range contributions.

$$B_{23} = B_{23}^{\mathrm{ex}} + B_{23}^{\mathrm{el}} + B_{23}^{sr} \tag{4.8}$$

The relationship between the excluded volume and short-range contributions to $B_2 3$ is determined by τ_{23}, as shown in Eq. 4.9, where τ_{23} corresponds to the strength of the adhesive force between the protein and the matrix.

$$B_{23}^{sr} = \frac{-B_{23}^{\mathrm{ex}}}{4\tau_{23}} \tag{4.9}$$

The electrostatic contribution to B_{23} is defined by the integral of the electrostatic potential between the protein and the matrix evaluated over the integration bounds of $r = R_h^a + R_h^b$ to $r = \infty$ to avoid integrating over the region of excluded volume between the two particles.

$$B_{23}^{\text{el}} = \frac{4\pi N_A}{M^a M^b} \int_{R_h^a + R_h^b}^{\infty} \left(1 - \mathrm{e}^{-\beta U(r)}\right) r^2 dr \tag{4.10}$$

The $\beta U(r)$ function, which describes the strength of electrostatic interaction between particles as a function of r, is described in Eq. 4.11:

$$\beta U\left(r\right) = Z^a Z^b \lambda_B \frac{\mathrm{e}^{-\kappa\left(r - R_h^a - R_h^b\right)}}{r\left(1 + \kappa R_h^a\right)\left(1 + \kappa R_h^b\right)} \tag{4.11}$$

The inverse Debye-Hückel length, κ is a measure of the distance that the electrostatic effect of a charge persists in a solution as a function of ionic strength, I.

$$\kappa^2 = 4\pi \lambda_B N_A I \tag{4.12}$$

The Bjerrum length, λ_B is the distance between two elementary charges at which the electrostatic interaction is equal to $k_B T$.

$$\lambda_B = \frac{\mathrm{e}^2}{4\pi \varepsilon_0 \varepsilon k_B T} \cong 0.7nm \tag{4.13}$$

In this model, we approximate the charge on each particle as a locally linear function of pH with parameters $a1$ and $a2$ as the slope and the intercept, respectively.

$$Z^a \cong a_1 pH + a_2 \tag{4.14}$$

A diagram of the equation relationships in this diffusion model is shown in Figure 4.1. 4.1 adapted a tree structure to illustrate how variables are passed on by equations explained above to calculate the diffusion coefficients D. The calculation process starts from the bottom layer of the tree (β,Z^a,Z^b,λ_B,κ), and works upwards by using equations represented by shaded equation numbers(e.g. represents $B_{23} = B_{23}^{ex} + B_{23}^{el} + B_{23}^{sr}$). The calculation process finishes when diffusion coefficients D is calculated , which folds in all values and relationships contained in the tree structure diagram, as illustrated in Figure 4.1.

4.2.2 Encoding Physics Level Parameters Into an ANN

Figure 4.2 is one simple example of an artificial neural network. It consists of an input layer, that matches the dimension of features of the experimental data; an intermediate layer, in which there are ten neurons; and lastly, an output layer that consists of just one neuron, which is also the dimension of the desired prediction space. Note that, if an ANN consists of more than one intermediate layer, then it is typically referred to as a deep neural network (DNN[22]). In the context of this chapter, the term DNN and ANN are loosely interchanged; as the structure of the ANN is not the key assignment to be solved, and the complexity of the ANN is far from "deep", i.e. instead of dozens of hidden layers, this chapter only explored into $2 \sim 3$ intermediate layers.

To incorporate the DVLO theory elucidated in section DLVO Theory with the classical ANN structure as shown in Figure 4.2, the full structure of the algorithm is shown in Figure 4.3. A first ANN is used for nonlinear regression between inputs (experimental conditions, diffusion constant recovered from the FRAP measurements detailed in section FRAP Theory and Experimental) and outputs (physics level parameters such as τ).

Input neurons are denoted *temperature*, *pH*, *normality*, which are color coded blue in Figure 4.3, meaning that they are either directly identifiable by looking up specs in databases, or they are independent variables that are directly quantifiable. The greed coded neuron, captioned *relaxation timescales*, is the independent variable from FRAP measurements.

The output layer consists of physics level parameters in the DLVO theory, in the simulations, we choose them to be *hydrodynamic radius*, *tau*, *charge* respectively. Additionally, *charge* is a secondary variable that is linked to a *slope* and *intercept* term in relation to *pH*, thus the notation of pH_A and pH_B. This output layer from the ANN is denominated as *Physics-Encoded Parameter Space*. The task of the ANN is to predict the designated physics parameters.

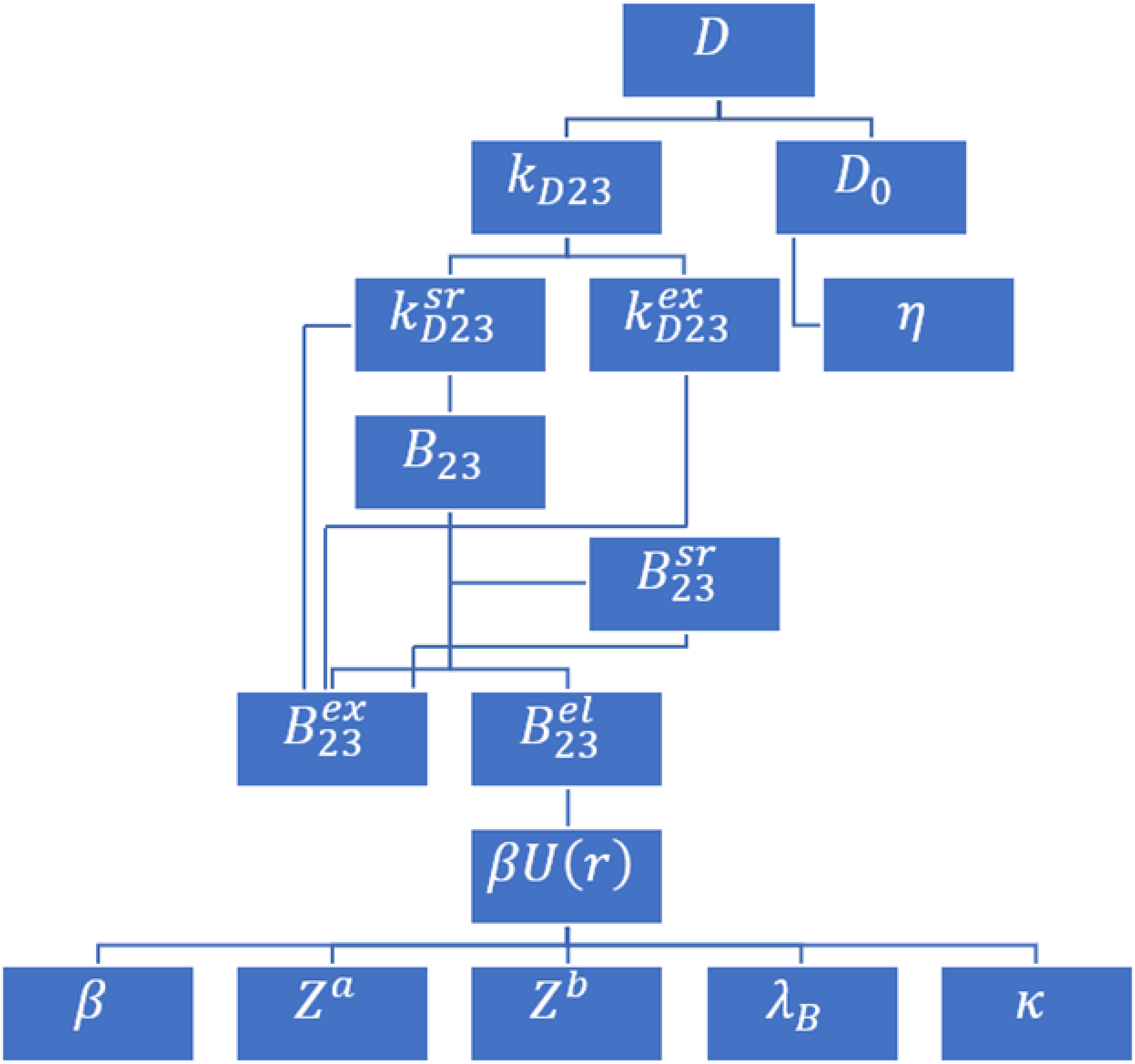


Figure 4.1. A graphical representation of the tree structure diagram. The diagram visualizes the tree structure relationship of variables and equations shown in theoretical framework. From the bottom layer (β,Z^a,Z^b,λ_B,κ) to the top layer (D), the tree diagram shows how equations connect variables to ultimately describe the measured diffusion coefficients D

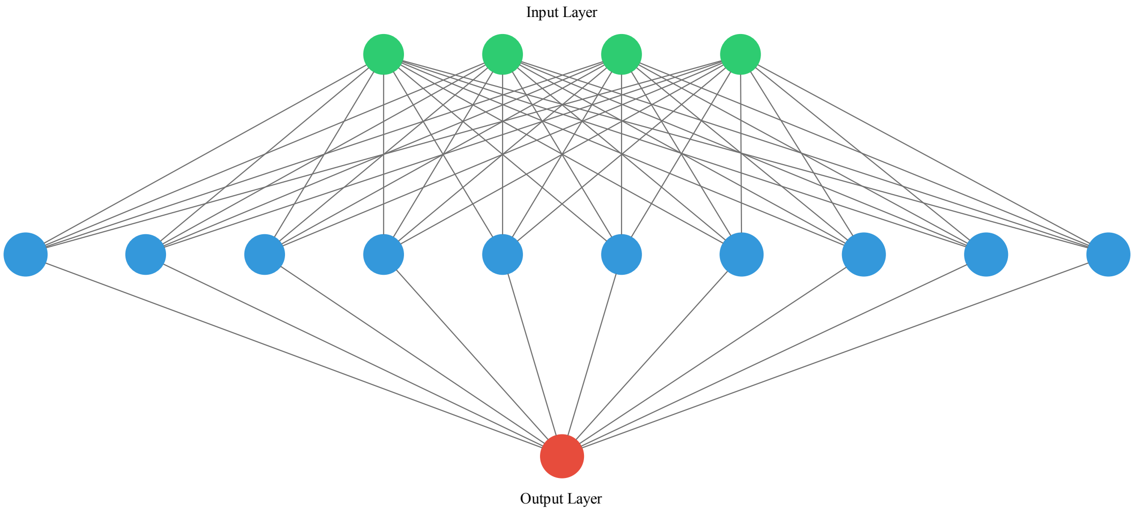


Figure 4.2. A simple example of Artificial Neural Network (ANN)

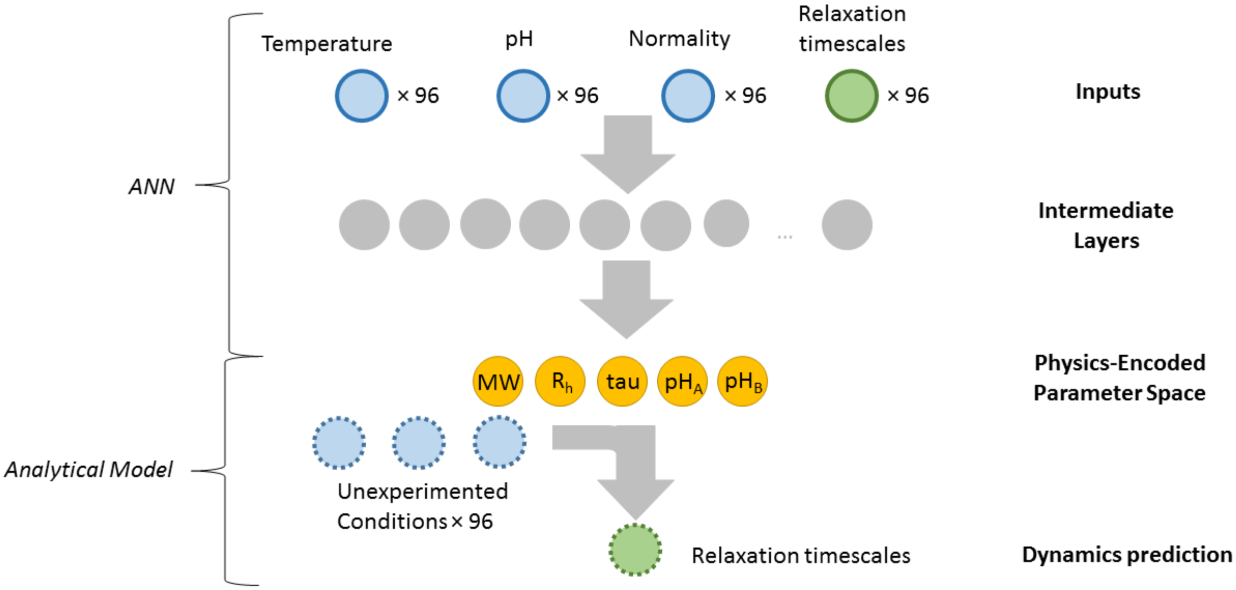


Figure 4.3. General structure of the algorithm. A first artificial neural network is used for nonlinear regression between inputs (experimental conditions, diffusion constant) and targets (latent space physical parameters). The diffusion constant in association with unmeasured experimental conditions are then calculated with the analytical model.

The second half of the algorithm is to use the analytical model described in DLVO Theory, to calculate the “relaxation timescales”. There is where this work and traditional ANN tasks differ. In a traditional setting, the task would be to train an ANN, such that the relaxation timescales can be predicted given the experimental design, i.e. this variable should locate at the output layer of the ANN. Instead, in this structure, relaxation time, even though it is a dependent variable, resides along with the other independent variables, together they serve as the input layer. The rationale behind this design is that, to train a traditional ANN to reach high accuracy prediction requires a lot of data. Furthermore, the ANN would still operate as a black box without offering any insight into the physical process.

4.2.3 FRAP Theory and Experimental

The experimental data were acquired using fluorescence recovery after photobleaching (FRAP). FRAP determines the kinetics of diffusion by bleaching a region of interest in the sample, and measuring how quickly fluorescent molecules move into the bleached area, thus providing information about the kinetics to the sample of interest[23].

Modelling the FRAP process and recovery a diffusion coefficient starting from Fick’s law of diffusion:

$$\frac{\partial}{\partial t}C(\rho,t) = \nabla\boldsymbol{D}\nabla\mathrm{C}(\rho,t) \tag{4.15}$$

Fick’s law of diffusion describes describes diffusive flux to the gradient of the concentration. The depth of field for the photobleaching laser is comparable to the overall thickness of $\sim$ 50 µm for the 700 nl droplets sandwiched between the slide and cover slip, such that the system can be reasonably assumed to be uniform in the z axis (i.e., for photobleaching with a pencil-beam). In this limit and for isotropic, normal diffusion, the general diffusion equation in Eq. 4.15 can be simplified to the following form:

$$C(\boldsymbol{r},t) = C(\boldsymbol{r},0) \otimes \left[(4\boldsymbol{\pi}\boldsymbol{D}t)^{-1}\,\mathrm{e}^{-\frac{\boldsymbol{r}^2}{4\boldsymbol{D}t}}\right] \tag{4.16}$$

In Eq. 4.16, D is the diffusion coefficient (now a scalar), t is time, $r(x,y)^2 = (x-x_0)^2 + (y-y_0)^2$ is the position vector within the lateral plane relative to the photobleaching origin, $C(r,t)$ is the time-dependent concentration profile, $C(r,0)$ is the initial concentration profile immediately following photobleaching, and the $\otimes$ symbol denotes convolution. According to Eq. 4.16, the concentration profile at $t > 0$ is given by the convolution of the initial concentration profile following photobleaching and a time-varying 2D Gaussian function.

If the initial photo bleached pattern is also reasonably described by a 2D Gaussian function imposed on a constant average concentration background of C_{ave}, the convolution of two 2D Gaussian functions has a relatively simple closed-form analytical solution (i.e., another 2D Gaussian function).

$$\begin{aligned} C(r,t) &= C_{ave}\left[1 - A\left(2\pi\sigma_b^2\right)^{-1} \mathrm{e}^{-\frac{r^2}{2\sigma_b^2}}\right] \otimes \left(4\pi D t\right)^{-1} \mathrm{e}^{-\frac{r^2}{4\boldsymbol{D}t}} \\ &= C_{\mathrm{ave}} - A\left(2\pi\sigma_b^2 + 4\pi\boldsymbol{D}t\right)^{-1} \mathrm{e}^{-\frac{r^2}{2\sigma_b^2+4\boldsymbol{D}t}} \end{aligned} \tag{4.17}$$

Leveraging radial symmetry, the position vector $\boldsymbol{r}$ has been replaced by a scalar value r of the displacement magnitude from the photo bleach origin. In addition, A is the relative depth of the photobleached spot ($A < 1$), and σ_b is the root mean squared deviation of the initial 2D Gaussian photobleach profile. This analysis implicitly assumes that the photobleach step arises over a timescale significantly faster than the subsequent recovery and can be modeled as instantaneous.

In the high-throughput FRAP analysis, a series of fluorescence images was acquired following photobleaching with a Gaussian excitation beam. The first post-bleach image was fit to a Gaussian, $G(x,y)$ of the functional form shown in Eq. 4.18, using a nonlinear least-squares fit with A, σ, x_0, and y_0 as adjustable parameters in the fit.

$$G(x,y) = I_{bkd} - A\mathrm{e}^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}} \tag{4.18}$$

I_{bkd} is the background fluorescence intensity from which the Gaussian bleach spot is subtracted. After x_0 and y_0 were obtained from the initial fit, each subsequent post-bleach image was fit to the same equation with a different nonlinear least-squares fit with only A and σ as adjustable parameters. After values for A and σ were obtained for the photobleached spot in each frame, the time dependent values of A were fit with Eq. 4.19 to recover the diffusion coefficient D and the recoverable fraction R.

$$A(t) = \frac{(A_0 - R)\,\sigma_0^2}{\sigma_0^2 + 2Dt} + R \tag{4.19}$$

A_0 and σ_0 are the best-fit values from the initial post-bleach frame ($t = 0$). Diffusion coefficient values were obtained for samples of varying protein and matrix composition, protein and matrix concentrations, ionic strength, and pH. The experimental conditions (pH, temperature T, ionic strength I, concentrations c), properties of the protein/matrix (hydrodynamic radius R_h, molecular weight M, charge behavior a_1 & a_2), and experimental results (diffusion coefficient D) were fit with the diffusion model described in the DLVO theory section to recover the adhesive force parameter, τ_{23} for each protein-matrix pair under investigation.

The traditional FRAP instrument comprises a typical microscope set-up, for which the experiments can laboratories and tedious, not to mention time-consuming when the images need to be analyzed frame by frame. We are using a high-throughput FRAP instrument from Formulatrix. Figure 4.4 is a schematic illustration of high-throughput FRAP instrumentation and data analysis. The FRAP instrument is capable of automatic analysis of 96 well plates prepared by robotic liquid handling. In a reasonably short amount of time ($\sim$ 4 hours), temperature, pH, salt concentration, composition, and concentration. The experimental conditions can be varied to obtain 96 measurements per plate of diffusion constants under highly reproducible conditions. The proteins are three different monoclonal antibodies (mAb) drug conjugates and BSA.

4.2.4 Data Simulation and Processing

First, we simulated FRAP data to validate the PENN architecture. Given the equation set described in DLVO Theory, random values following a uniform distributions were gener-

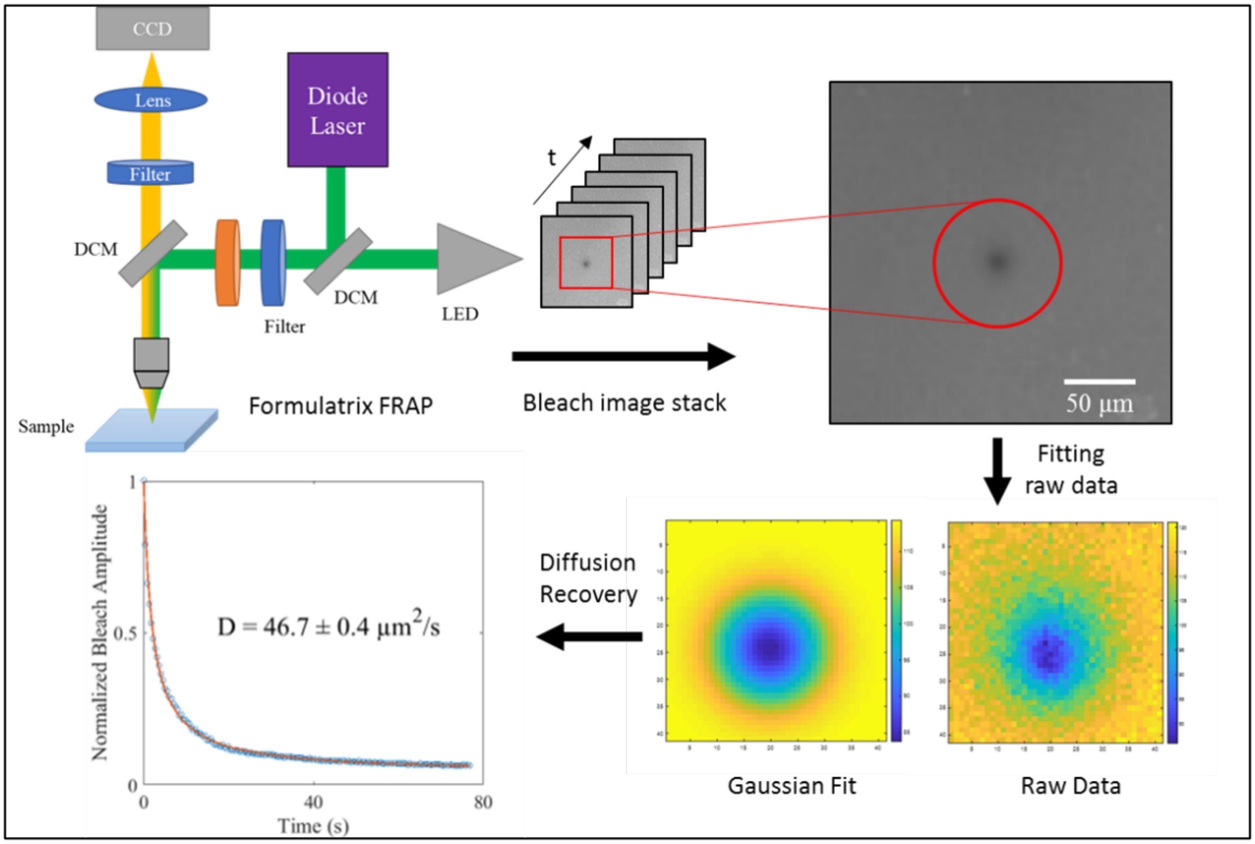


Figure 4.4. Schematic illustration of high-throughput FRAP instrumentation and data analysis overview. DCM: dichroic mirror. Protein samples were placed on 96 wells plate and FRAP image stack was collected in real-time. Region of interest was crossed out to do least square. The convolution of time dependent 2D Gaussian with initial concentration gave time-dependent concentration, followed by FRAP analysis to recover diffusion coefficients.

ated for each of the variables involved in the equations. The data simulation is summarized in Table 4.1.

Table 4.1.
Data Simulations

	physics definition	range low	range high	unit
pH	pH	5	9	
T	temperature	293.15	305.15	K
N	normality	0.01	0.05	M
M	molar mass	125e3	175e3	Dalton
Ph	hydrodynamic radius	4.5	6.5	nm
a1	offset, Z versus pH plot	50	100	
a2	slope, Z versus pH plot	-25	25	
tau	“stickiness“	0.1	2	

And the involved constants are listed in Table 4.2

To generate simulated data, two additional parameters are also defined: p and e, representing types of proteins and the repetitive experiment for per protein respectively. Note that, for the simulated data, only protein-protein interaction is considered. Additionally, a standard scalar is applied to all data such that the ANN is considering numbers on the same scale. Everything is evaluated upon a reverse transformation.

Table 4.2.
Constants in DLVO

symbol	physics definision	value	unit
kB	Boltzmann constant	1.3807e-16	$\mathrm{g\,cm^2\,s^{-2}\,K^{-1}}$
N_A	Avogadro’s number	6.0221409e23	K
λ_B	Bjerrum length	0.7	nm
A	viscosity constant A (water)	2.414e-4	$\mathrm{g\,cm^{-1}\,s^{-1}}$
B	viscosity constant B (water)	247.8	K
C	viscosity constant C (water)	140	K

For the experimental data that were acquired using the FRAP instrument as described in section FRAP Theory and Experimental, the model was extended to a protein-matrix interaction, which involves a B_{23} term instead of B_{22}. For the simulated data set, T, pH, N, D were stacked together as the ANN input, *a1*, *a2*, M, *Rh*, *tau* were stacked together as the ANN output, i.e. to be predicted by the ANN. For the FRAP experimental data, only *tau* were considered output and everything else were considered input. The rationale will be discussed in the Results and Discussion section. The ANN model was built using Keras[24], *Adam optimizer* was used for the training optimization, loss function was chosen to be *mean squared error*, and the metrics used during the training was *accuracy*.

4.3 Results and Discussion

4.3.1 PENN With Simulated Data

PENN combined a simple black-box ANN and an analytical model, instead of a straightforward traditional ANN. The rational for this design is that, the analytical model offers physics constraints such that ANN can be trained with less data. As is known, experimental data are expensive to acquire, even more so if desiring to reach the amount for "big data analysis". Therefore, in consideration of the limited data size part of the uncertainty is reduced by imposing mathematical connections within the variable.

Consequently, the variables produced at the output layer of the ANN exhibit high covariance because of the mathematical connections within. It is hypothesized that they reside on a hyper-surface. The task of recovering these variables would be analogous to fitting a curve to a high-order polynomial. The quality of the fit can be improved by adding more terms in the polynomial expansion, but at the expense of higher covariance and lower statistical confidence in the recovered parameters. In other words, the result of the regression analysis will not be a unique set of parameters, but rather a "feature" (hyper-surface) within parameter space defined by multiple combinations of parameters of comparable probability.

Using an initial 1000 types of protein, per protein-protein simulated data was repeated 100 times, with 100 epochs and a batch size of 20, The training of the model reports accuracy and loss as shown in Figure 4.5. Note that, the train-validation split is separate from the testing dataset to be used later for the analytical model, i.e. the train-test split shown in Figure 4.5 is a subset of the training data that was allocated for ANN. To reiterate, the dataset is first split into *train* and *test*, when *train*was fed into the ANN, upon fitting the model, *train* was further split into *train* and *val*; and *test* was reserved for testing independent of the ANN training process. In other words, *train* is the sub-dataset used for training the ANN; *val* is the sub-dataset used to determine whether the ANN over-fits or under-fits; and *test* is the sub-dataset used to determine the accuracy of the entire model, ANN included. The nomenclature for *train val* and *test* here slightly differs from when they are used in cross-validation.

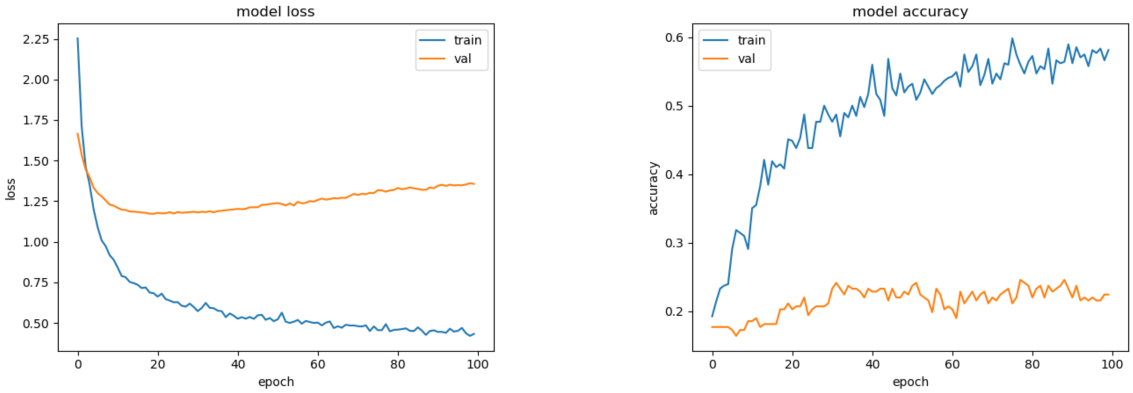


Figure 4.5. Plot of model accuracy and loss during the training process

Upon inspection of Figure 4.5, the line indicating the training data reaches relative stable accuracy at about 60 epochs, when the loss for training data is low. Yet the accuracy for the training is not ideal, at ~ 0.6. And this model doesn't generalize well to the testing data, as the orange curve doesn't comply with the blue curve. The trends indicate that, this model is not well-trained to be applied to reliably predict the data, which is confirmed with *test* dataset as displayed in Figure 4.6.

Traditionally, this is where the design of the ANN be improved such that higher accuracy can be achieved. Numerous actions can be taken to achieve this goal. For example, increasing the number of the hidden layer, altering the number of neurons per layer, try different batch

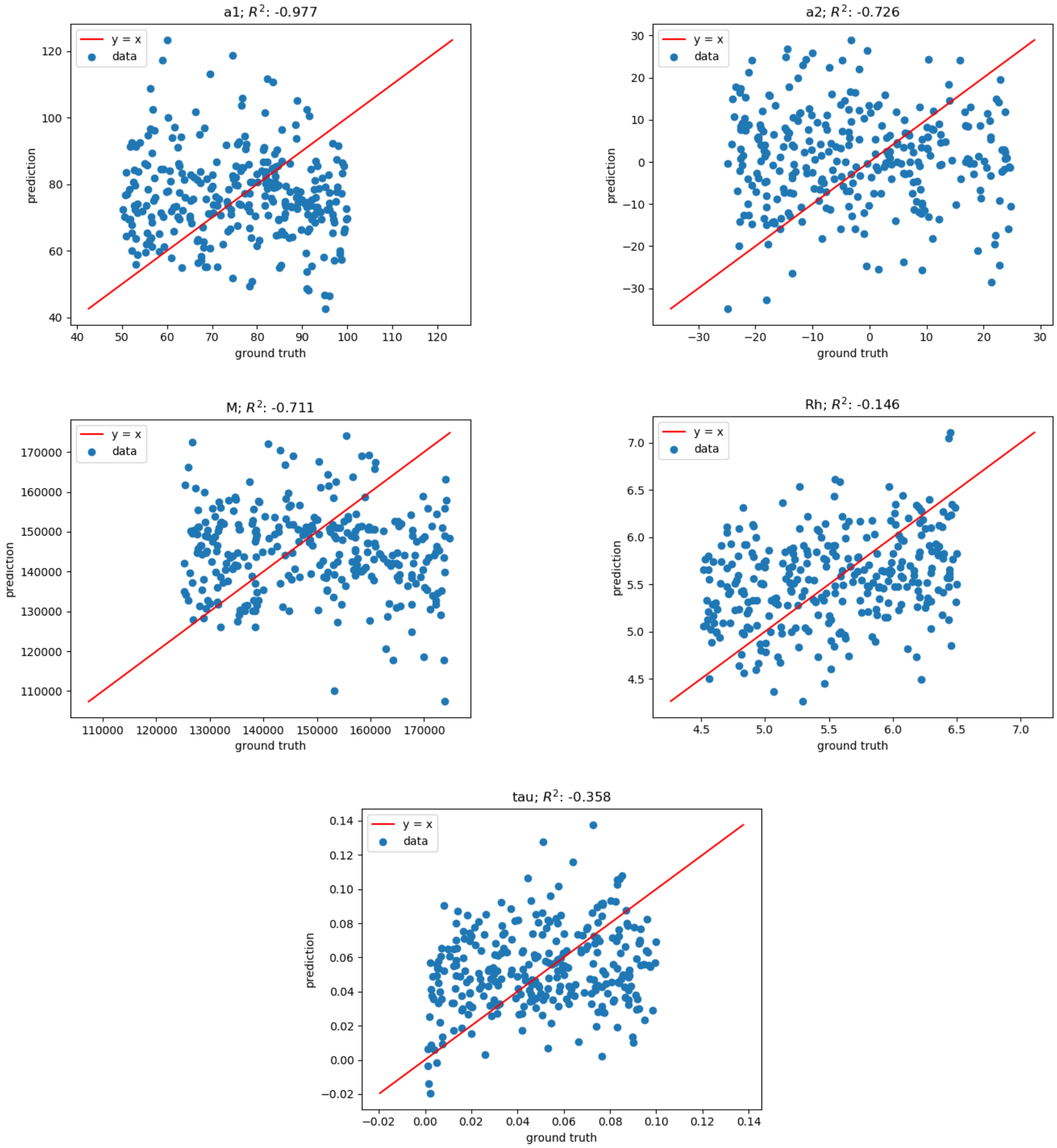


Figure 4.6. ANN prediction of the physics parameters

sizes, try different activation functions, etc. Essentially, there is an unlimited number of hyper-parameters can be tuned even though with diminishing result.

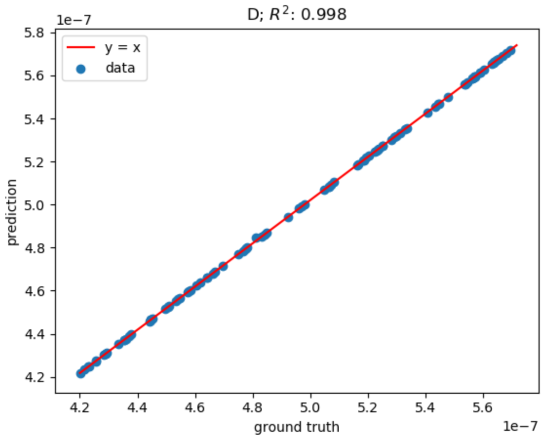


Figure 4.7. Diffusion coefficient recovered by PENN using simulated data.

However, when we took the output of the ANN, to calculate *diffusion coefficient* using the analytical model, a nice agreement between the ground-truth and the recovered *diffusion coefficient* was achieved, the result is shown in Figure 4.7. This agreement supports the hypothesis that, when using an ANN as a tool for regression analysis, the uncertainty shared by the physics layer variables, cancels each other out after combining using the analytical model, i.e. uncertainty for the physics layer variables reside on a hyper-surface.

4.3.2 Least Squares Fit Recovered Ground Truth for Experimental Data

The concept of PENN has been successfully demonstrated to recover the diffusion constant as shown in the previous Section(4.3.1) to practically meaningful FRAP data, it is first important to have the ground truth. In the simulations, ground truth of most variables are randomly generated, assuming a known range (except for diffusion constant for a given protein-protein pair, which was calculated using the DLVO model).

However, getting the ground truth of the experimental data required an additional step, as the *tau* (τ) parameter is not an experimental measurable.

Given the known variables summarized from databases as shown in Table 4.3 (In the charge behavior column, a is the slope and b is the y-intercept of the Z vs. pH linear fit), and D (diffusion constant) is an output from the FRAP image analysis, therefore, known. The equations in section DLVO Theory combined serves as an implicit function of τ, which can be solved via a least-square fit, assuming τ is a constant given a protein-matrix combination pair.

Table 4.3.
Protein/matrix variables in the DLVO model.

Parameter	Molecular weight	Hydrodynamic radius	Charge behavior
Method	SEC	SEC	Matrix - zeta potential Protein - amino acid pKa
HA	1500 kDa	81 nm	a = -194 e/pH; b = 100 e
Collagen	280 kDa	50 nm	a = -34 e/pH; b = 249 e
BSA	66.5 kDa	2.68 nm	a = -15 e/pH; b = 93 e
mAb 1	149 kDa	3.51 nm	a = -19 e/pH; b = 163 e
mAb 2	146 kDa	3.48 nm	a = -19 e/pH; b = 163 e
mAb 3	150 kDa	3.51 nm	a = -19 e/pH; b = 163 e

Setting the fitting parameters as:

```
tau_ini = 10 * rand(2, 4);
options = optimoptions('lsqnonlin', 'StepTolerance',1e-20, 'FunctionTolerance', 1e-20);
options.Algorithm = 'levenberg-marquardt';
```

Recovered τ values as shown in Table 4.4.

Before setting the τ values listed in Table 4.4 as the ground truth for ANN, it's useful to estimate the uncertainties in those values. The major sources of uncertainties come from i) the physics constants in Table 4.3, as they represent the average of the selected analyte, not necessarily the actual experimental values (summarized in Table 4.5); ii) the uncertainties in experimental variables, when the intended deviated from the actual, such as the concentration of the matrices, which is hard to be accurate, to begin with; iii) the uncertainty from the least-squares fit.

Table 4.4.
τ recovered from least squares fit

	mAb1	mAb2	mAb3	BSA
HA	1.31	1.40	8.89	2.41
Collagen	0.87	0.69	1.37	1.16

Table 4.5.
Uncertainty values for parameters in the DLVO model.

Parameter	Measure of uncertainty	RSD
Protein/matrix concentration, c	Liquid handler specs	3.50%
Matrix molecular weight, Mb	SEC	1%
Matrix hydrodynamic radius, Ma	SEC	1%
Protein hydrodynamic radius, Rha	SEC	0.50%
Temperature, T	Temperature logger	0.01%
Charge behavior, Z	Zeta potential/amino acid charge	Error from fit

To account for all of the uncertainties mentioned above, the error propagation equation based on Taylor linearization is used to estimate the uncertainties in τ.

As τ is implicitly given by $\psi\left(\boldsymbol{x_{\mathrm{i}}}, D\right) = \tau$, with $\boldsymbol{x_{\mathrm{i}}} \in \{experimental variables\}$. Therefore, uncertainty in τ can be calculated as shown is Eq. 4.20:

$$\sigma_{\tau}^{2} = \sigma_{fit}^{2} + \sum_{x_{\mathrm{i}}} \left[\left. \frac{\partial \psi\left(u_{\mathrm{i}}\right)}{\partial x_{\mathrm{i}}} \right|_{u_{\mathrm{i}}} \right]^{2} \sigma_{x_{\mathrm{i}}}^{2} \tag{4.20}$$

Uncertainty from the fit was obtained from the covariance matrix, $\mathbf{X}$ which was calculated from the Jacobian, $\mathbf{J}$ and the variance in the residuals, v_R, as shown in Eq. 4.21. The variance in each fitting parameter, σ^2_{fit} is found along the diagonal of the covariance matrix, $\mathbf{X}$[25].

$$\mathbf{X} \; = \; v_R \left(\mathbf{J}^T \times \mathbf{J} \right)^{-1} \tag{4.21}$$

As such, the uncertainties associated with τ is compiled in Table 4.6:

From Table 4.6, we conclude that the τ value is relatively accurate, in which the general uncertainty is about 10 percent. Bearing in mind that, *uncertainty* becomes less relevant once data was put into the ANN. Therefore Table 4.6 serves more purpose when discussing ways to improve later on in the chapter.

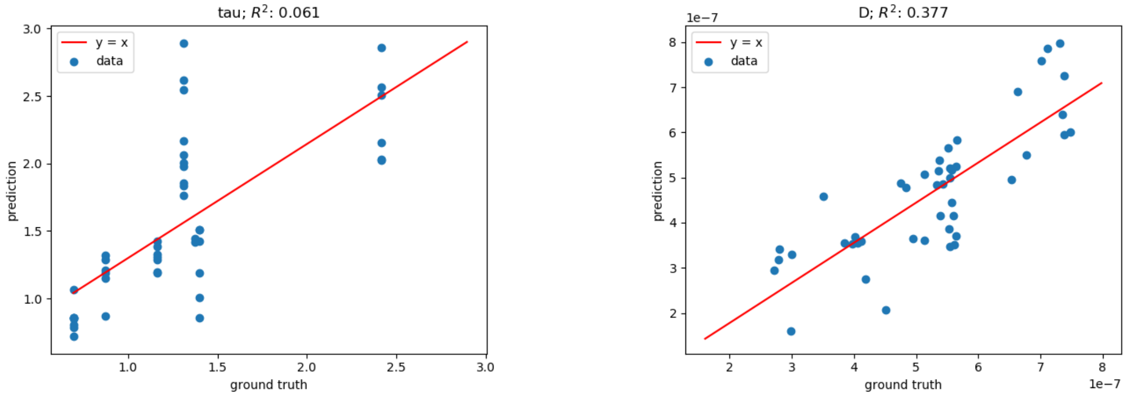
Table 4.6.
Uncertainties in prediction of τ calculated via error propagation.

	mAb1		mAb2		mAb3		BSA	
	mean	SD	mean	SD	mean	SD	mean	SD
HA	1.31	0.11	1.39	0.11	8.89	9.04	2.41	0.32
Collagen	0.87	0.11	0.69	0.06	1.37	0.60	1.16	0.11

4.3.3 PENN With Experimental Data

In the previous section (4.3.2Least Squares Fit Recovered Ground Truth for Experimental Data), the ground-truth was calculated using the physics parameter τ. The PENN used in section PENN With Simulated Data was then modified, such that only τ was the *Output* of the ANN, and the rest of the variables were stacked up as the *Input*. With the same ANN parameters as in the simulation, experimental data produced results as shown in Figure 4.8.

Comparing Figure 4.8a with 4.6 (the ANN output for simulated data), given that the $y = x$ line is what the ANN model is trying to regress to, all of them have low correlation coefficient R^2. From Figure 4.8a to Figure 4.8b, R^2 improved significantly. This phenomenon is expected as explained before: the physical parameters are correlated to one another.



(a) ANN prediction of the physics parameter τ.

(b) Diffusion coefficient recovered by PENN.

Figure 4.8. Applicate PENN to FRAP experimental data.

Figure 4.8b illustrates the final result produced by PENN using the experimental FRAP dataset. The regression coefficient R^2 was calculated to be 0.377, which is not as optimal as the simulated dataset as shown in Figure 4.7, Nevertheless, a clear trend where the data-points gravitate towards the $y = x$ line is observed, suggesting that the concept of incorporating physics parameters into ANN improves the predicting accuracy compared to that from the raw ANN output.

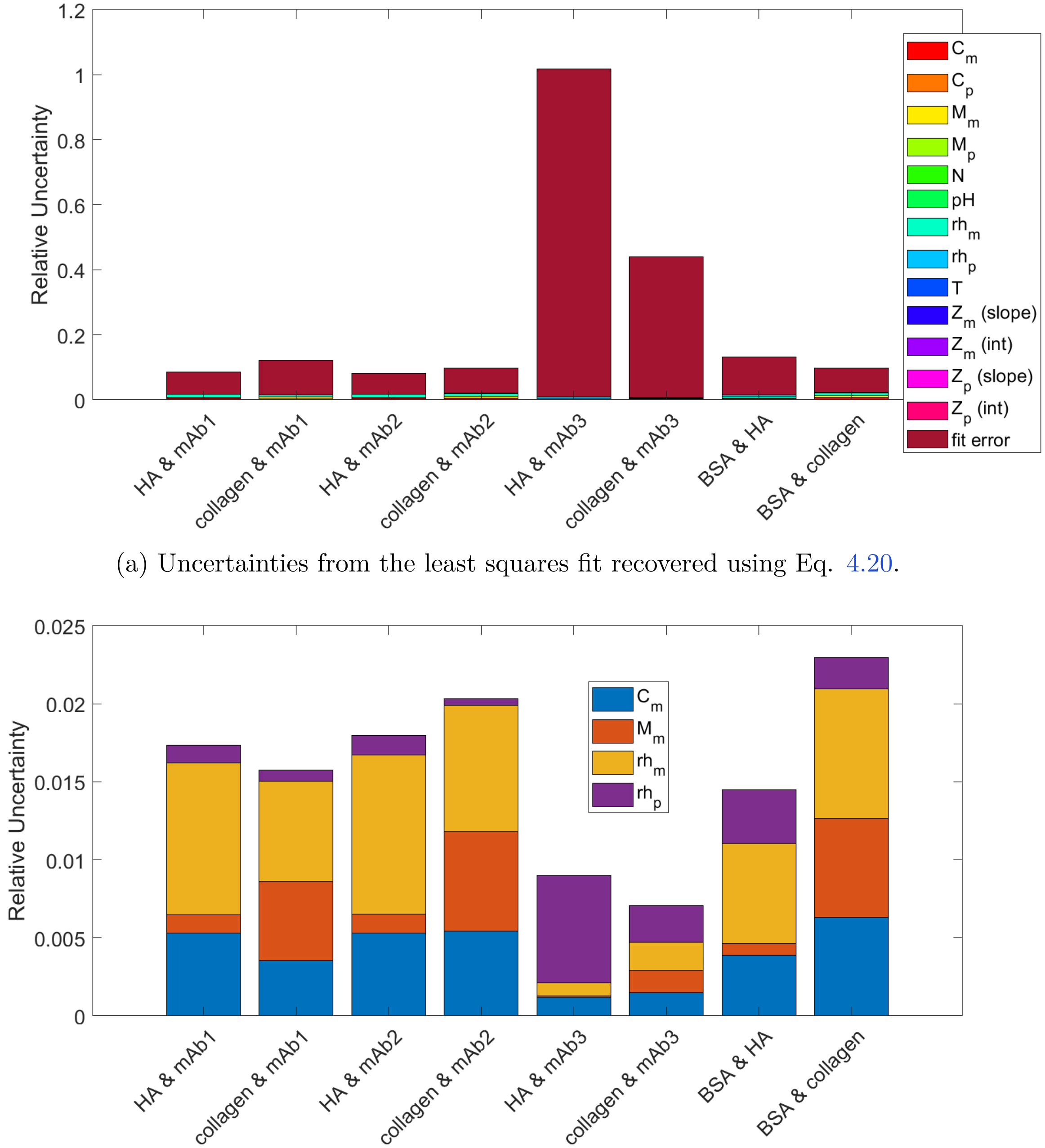
4.3.4 Uncertainty Evaluation and Future Developments

A detailed breakdown of uncertainties tabulated in Table 4.6 is visualized in Figure 4.9. As shown in 4.9a, the most uncertainty comes from the least-squares fit. During the fitting process, the tolerance had been set to a minimum of 1.0×10^{-20}. Because the least-squares method minimizes the sum of the offsets/residuals of data points, most likely, the uncertainties from the fit arise from instrumentation systematic error. For example, from Table 4.5, the reported RSD of the liquid handler is 3.5% from the manufacturer's website; realistically we observe variances from drop to drop in the 96 well-plate to be visually noticeable thus the protein/matrix concentration might not be accurate.

Upon close inspection of Eq. 4.1, the concentrations are on the first order of the expansion series, therefore its' accuracy carries greater impact of the result's accuracy, i.e. if the factor in the expansion/additional deviations, as whatever calculated with the diffusion coefficient k is multiplied by the concentration.

Alternatively, the concentration could be considered an unknown variable be recovered through least-squares fit. In which case it would be a two-parameter (τ plus concentration) fit instead of one-parameter (τ only). However, the equations in 4.2.1 involve integration, therefore, nonlinear, it would need many repeated trials of the same condition for the recovery, which would be experimentally expensive, and against the design of PENN, i.e. using mathematical relations to account for the lack of training data.

Future development can also consider exchanging the contents of what are currently in *Inputs* and *Outputs* in the ANN, as long as *Diffusion Constant* remains in the *Inputs*, the design of the PENN is conserved. As much as this is opening up the possibilities of



(a) Uncertainties from the least squares fit recovered using Eq. 4.20.

(b) Uncertainties from selective experimental variables, the variables were selected because they account for the majority (> 95%) of the experimental variance.

Figure 4.9. Uncertainties contributions from different factors

the architecture, it is also increasing the number of hyper-parameters to be tuned, which increases the complexity of the architecture.

Finally, currently, the ANN contains a simple one hidden layer of ten neurons, previously it has been rationalized that, with the lack of complexity of the feature space of the FRAP data, there's diminishing returns trying the increase either the number of hidden layers or the number of neurons for increasing the accuracy of the ANN. However, should we choose to reorder what are stacked in *Inputs* and *Outputs* respectively, it will be worth looking into increase the complexity of the ANN.

4.4 Conclusions

This chapter described the concept of incorporating physics level parameters into artificial neural networks. The physics level parameters residing at the *Outputs* layer are mathematically related, such that the individual prediction of each is not necessarily accurate, but the final prediction of PENN is of much higher accuracy than that of the intermediate layer. The architecture was proven successful with a simulated dataset, and validated by actual dataset acquired using FRAP. The bottleneck to improve the predicting accuracy further lies in eliminating systematic error from the instrumentation, which could potentially be difficult to eliminate, alternative algorithmic improvement was discussed as well. PENN is a promising algorithm to greatly reduce the effort to gain access to the diffusion constant avoiding the bench-top measurement. It could potentially offer insight into the protein interactions theory.

References

[1] M. J. Rathbone, J. Hadgraft, M. S. Roberts, and M. E. Lane, *Modified-release drug delivery technology, second edition*. 2008, vol. 1.

[2] R. G. Thorne, S. Hrabětová, C. Nicholson, M. Stroh, and W. M. Saltzman, *Diffusion measurements for drug design (multiple letters)*, 2005. DOI: 10.1038/nmat1489. [Online]. Available: www.nature.com/naturematerials.

[3] Y. Wang, C. Li, and G. J. Pielak, "Effects of proteins on protein diffusion," *Journal of the American Chemical Society*, vol. 132, no. 27, 2010, ISSN: 00027863. DOI: 10.1021/ja102296k.

[4] F. Roosen-Runge, M. Hennig, F. Zhang, R. M. Jacobs, M. Sztucki, H. Schober, T. Seydel, and F. Schreiber, "Protein self-diffusion in crowded solutions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 29, 2011, ISSN: 00278424. DOI: 10.1073/pnas.1107287108.

[5] J. T. Loessberg-Zahl, M. Gillrie, R. D. Kamm, A. van den Berg, A. van der Meer, and J. C. Eijkel, "Diffusion from steady-state profile (DSSP) for low cost, low concentration measurement of diffusion," in *23rd International Conference on Miniaturized Systems for Chemistry and Life Sciences, MicroTAS 2019*, 2019.

[6] M. H. Hettiaratchi, A. Schudel, T. Rouse, A. J. García, S. N. Thomas, R. E. Guldberg, and T. C. McDevitt, "A rapid method for determining protein diffusion through hydrogels for regenerative medicine applications," *APL Bioengineering*, vol. 2, no. 2, 2018, ISSN: 24732877. DOI: 10.1063/1.4999925.

[7] G. Frenning and M. Strømme, "Drug release modeled by dissolution, diffusion, and immobilization," *International Journal of Pharmaceutics*, vol. 250, no. 1, pp. 137–145, Jan. 2003, ISSN: 03785173. DOI: 10.1016/S0378-5173(02)00539-2.

[8] A. M. Thomson and J. L. Perry, *Collaboration processes: Inside the black box*, 2006. DOI: 10.1111/j.1540-6210.2006.00663.x.

[9] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, 2018, ISSN: 21693536. DOI: 10.1109/ACCESS.2018.2870052.

[10] J. A. Tulsky, G. S. Fischer, M. R. Rose, and R. M. Arnold, "Opening the black box: How do physicians communicate about advance directives?" *Annals of Internal Medicine*, vol. 129, no. 6, 1998, ISSN: 00034819. DOI: 10.7326/0003-4819-129-6-199809150-00003.

[11] R. Chen, X. Liu, S. Jin, J. Lin, and J. Liu, "molecules Machine Learning for Drug-Target Interaction Prediction," DOI: 10.3390/molecules23092208. [Online]. Available: www.mdpi.com/journal/molecules.

[12] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, Feb. 2019, ISSN: 10902716. DOI: 10.1016/j.jcp.2018.10.045.

[13] L. Sun, H. Gao, S. Pan, and J. X. Wang, "Surrogate modeling for fluid flows based on physics-constrained deep learning without simulation data," *Computer Methods in Applied Mechanics and Engineering*, vol. 361, p. 112 732, Apr. 2020, ISSN: 00457825. DOI: 10.1016/j.cma.2019.112732. arXiv: 1906.02382.

[14] Y. Zhu, N. Zabaras, P. S. Koutsourelakis, and P. Perdikaris, "Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data," *Journal of Computational Physics*, vol. 394, pp. 56–81, Oct. 2019, ISSN: 10902716. DOI: 10.1016/j.jcp.2019.05.024. arXiv: 1901.06314.

[15] C. Rao, H. Sun, and Y. Liu, "Physics-informed deep learning for incompressible laminar flows," *Theoretical and Applied Mechanics Letters*, vol. 10, no. 3, pp. 207–212, Mar. 2020, ISSN: 20950349. DOI: 10.1016/j.taml.2020.01.039. arXiv: 2002.10558.

[16] G. Kissas, Y. Yang, E. Hwuang, W. R. Witschey, J. A. Detre, and P. Perdikaris, "Machine learning in cardiovascular flows modeling: Predicting arterial blood pressure from non-invasive 4D flow MRI data using physics-informed neural networks," *Computer Methods in Applied Mechanics and Engineering*, vol. 358, p. 112 623, Jan. 2020, ISSN: 00457825. DOI: 10.1016/j.cma.2019.112623. arXiv: 1905.04817.

[17] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, "Parameterized Machine Learning for High-Energy Physics," Tech. Rep., 2016. arXiv: 1601.07913v1.

[18] M. Hermansson, "The DLVO theory in microbial adhesion," *Colloids and Surfaces B: Biointerfaces*, vol. 14, no. 1-4, pp. 105–119, 1999, ISSN: 09277765. DOI: 10.1016/S0927-7765(99)00029-6.

[19] B. W. Ninham, *On progress in forces since the DLVO theory*, 1999. DOI: 10.1016/S0001-8686(99)00008-1.

[20] D. Roberts, R. Keeling, M. Tracka, C. F. Van Der Walle, S. Uddin, J. Warwicker, and R. Curtis, "The role of electrostatics in protein-protein interactions of a monoclonal antibody," *Molecular Pharmaceutics*, vol. 11, no. 7, pp. 2475–2489, Jul. 2014, ISSN: 15438392. DOI: 10.1021/mp5002334.

[21] D. N. Petsev and N. D. Denkov, "Diffusion of charged colloidal particles at low volume fraction: Theoretical model and light scattering experiments," *Journal of Colloid And Interface Science*, vol. 149, no. 2, 1992, ISSN: 00219797. DOI: 10.1016/0021-9797(92)90424-K.

[22] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017, ISSN: 18728286. DOI: 10.1016/j.neucom.2016.12.038.

[23] D. Axelrod, D. E. Koppel, J. Schlessinger, E. Elson, and W. W. Webb, “Mobility measurement by analysis of fluorescence photobleaching recovery kinetics,” *Biophysical Journal*, vol. 16, no. 9, pp. 1055–1069, 1976, ISSN: 00063495. DOI: 10.1016/S0006-3495(76)85755-4. [Online]. Available: /pmc/articles/PMC1334945/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1334945/.

[24] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems*. 2019, p. 851, ISBN: 9781492032649. [Online]. Available: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/.

[25] S. Nocedal, Jorge and Wright, *Numerical optimization*. Springer Science & Business Media.

5. HYPERSPECTRAL INFRARED IMAGING

5.1 Introduction

In accelerated stability testing of active pharmaceutical ingredients (APIs), a critical step in the formulation development pipeline, APIs are exposed to elevated humidity and temperature ranges for extended periods of time to mimic the storage conditions, such that the stability of the API can be determined. If crystalline APIs transition from one polymorph to another under these conditions and the polymorphs exhibit varying degrees of solubility, and consequently, bioavailability medication may lose its potency or have unanticipated side-effects, endangering patients. Even slight changes in the overall composition of the API can dramatically affect patient outcomes. The standard assay techniques to test for API crystallization include titration, capillary electrophoresis, chromatography, spectroscopy, and electroanalytical methods[1]. Of these methods, optical spectroscopic methods carry the advantages of non-invasive, therefore suitable for fast screening processes. Vibrational microscopy, especially infrared (IR) imaging, yields a complete IR spectrum at each pixel, provides chemical sensitivity, allows identification of target moieties across heterogeneous systems, and enables label-free characterization of active pharmaceutical ingredients. Due to the varied applications, IR imaging is one of the most widely adopted analytical methods to date. Infrared spectroscopy has a rich history in molecular characterization, leading to well-established libraries for various molecules and different conformations, as well as wide applications in pharmaceutical[2], biomedical[3] and metrology applications[4], thus serving as a potential candidate for real-time monitoring of stability monitoring of the APIs of interest.

To achieve real-time monitoring, the drug production pipeline calls for flexible, versatile imaging modalities, especially for the drugs existing in the form of crystals capable of presenting multiple polymorphs[5]. As such, when analyzing these samples, it is advantages to evaluate structure on a per-particle basis, preferably through a real-time and non-invasive technique. IR imaging with quantum cascade lasers (QCLs) is an option for an imaging modality that can provide advantages over current benchtop techniques. Unlike common light sources (low-brightness broadband thermal or synchrotron sources) fast widely tunable

QCLs afford potential real-time IR imaging, on top of being more portable. Due to a focus on spectral ranges and packaged QCL sources (i.e., the usability for QCLs[6]), since the initial demonstration of QCLs[7], efforts have mainly been made in combining IR microscopy and QCLs.

A parallel effort of integrating QCLs into microscopes is in the exploration of classifiers for hyperspectral classification upon acquiring the information-rich IR spectra. Kuepper *et al.* took advantage of the tunability of QCLs to achieve short acquisition time for high-quality diagnostic images[8], in which they also used a simple random forest (RF) classifier for fast classification of the stained brain tumor tissue sections, this is the first demonstration of clinical incorporation, which proved QCLs to be a reliable substitution for FTIR-based microscopes in the clinical environment. Similarly, Yeh *et al.* optimized the configuration such that the signal-to-noise ratio is comparable to the fastest available HS FT-IT imaging system[9]. Other similar systems have been constructed with minor variations in the design, with occasional integration of ML methods.

In this chapter, a design to enhance spatial resolution of IR microscopy is introduced, leveraging both the tunability of QCLs and the integration of machine learning tools. IR transmittance images were acquired at 1190–1340 nm^{-1} and visible images of 640 nm laser transmittance of spherical particles were acquired simultaneously. IR and visible images were registered via cross-correlation USAF-1951 test grid images collected for both wavelength regimes. Following image registration, chemical classification was conducted via a single-layer neural network on each pixel within the IR images. Image segmentation conducted via image morphological analysis enabled the selection of individual particles within the visible image. A consensus chemical classification for pixels within each particle was then utilized to determine particle chemical identity, which was then visualized by color mapping the original visible-intensity image. In this manner, chemical sensitivity from IR spectroscopy was encoded into the visual images with resolution on the order of the visible wavelength, increasing an order of magnitude gain in theoretical resolution over direct imaging with IR wavelengths. This process was first applied to a test system of glass beads and poly-methyl-methacrylate beads on the order of 50 µm in diameter, and second applied to distinguish the Form I and Form II polymorphs of clopidogrel bisulfate, an inhibitor of blood platelet aggregation. The molecular formation of polymorph form I is known to be chemically active for its intended treatment, while form II is inactive, as it exhibits slower dissolution, kinetics and corresponding reductions in bioavailability. The detection between the two species is synonymous with that of a polymorph transition for the drug real-time monitoring pipeline.

5.2 Methods

5.2.1 Microscope Apparatus

The microscope is constructed as shown in Figure 5.1. A 640 nm diode laser producing the visible beam was combined with an IR beam, emitted by the QCL. Prior to the combination, the 640 diode laser beam was 4f coupled to allow for fine adjustment of the collimation to match that of the IR beam at the germanium window. The two co-propagating beams were expanded twice with two sets of 4f lenses and scanned across the sample using a pair of galvanometer mirrors, then after beam expansion, they were focused with a Cassegrain objective onto the sample plane. Light from the sample plane was re-collimated using a collection lens, and a germanium window was used to separate the 640 nm beam from the IR beam. The 640 nm light was detected using a photo-multiplier tube (Hamamatsu). IR light was detected using an IR-sensitive photodiode (Vigo Systems).

Responses of the detectors were digitized synchronously with a 10 MHz clock using a digital oscilloscope card (ATS9350, AlazarTech). Custom software (MATLAB) was used to down-sample raw data to coincide with IR laser pulses and remap the down-sampled data onto a set of 256 pixel × 256 pixel IR images (one for each QCL channel) and one 256 pixel × 256 pixel 640 nm bright-field image.

5.2.2 Sample Preparation

Dichloromethane (DCM) was purchased from Sigma-Aldrich. To prepare the mixture of DCM and water, they were mixed in a sealed tube, upon vigorous shaking, the mixture were pipetted onto a CaF_2 slide. Polymethyl methacrylate (PMMA) and silica beads were purchased from Cospheric. Clopidogrel bisulfate form I and form II were produced in-house at Dr. Reddy's Laboratories and were used as received. For these solid beads samples,

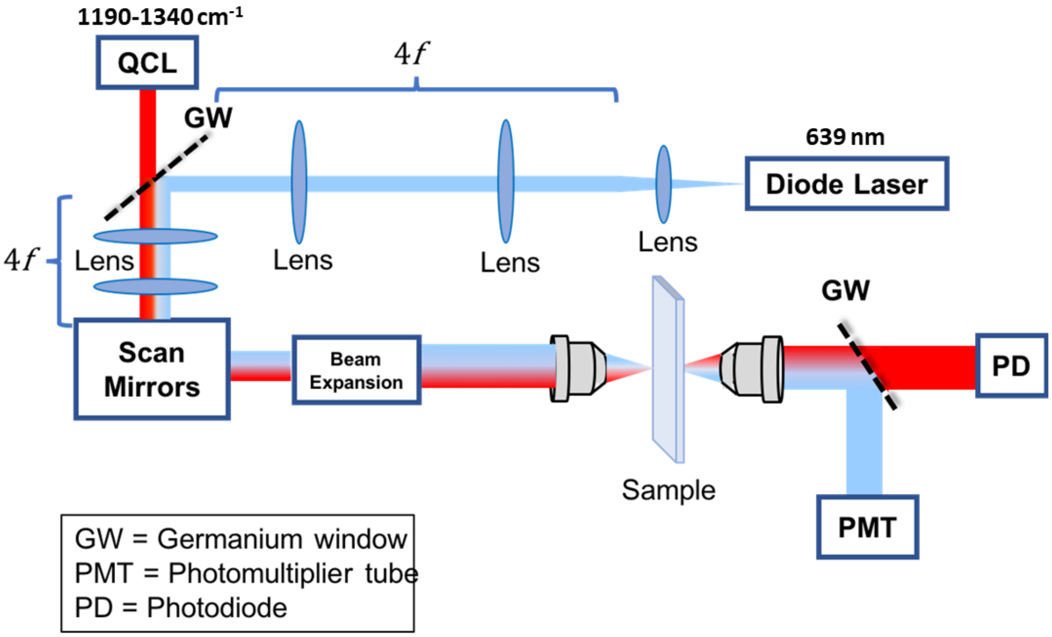


Figure 5.1. Instrument schematic of the hyperspectral microscope.

training data for particle classification were acquired from images of pure components; testing data were acquired from images with equally mixed beads.

5.2.3 Digital Data Processing

Digitization was initialized by a function generator that triggers both the Alazar digital acquisition card and the QCL. A custom MATLAB script was used to acquire 6240 digitization events at a rate of 160 MHz, corresponding to an acquisition duration of 40 µs. Prior to any experiments, a blank image was acquired then taken the first derivative to determine the rise and edges of each pulse, resulting in an integration matrix, i.e., a mask to apply to each sample to obtain signal only where the laser is firing. The integration matrix has a conditionality of n × 6240, n being the active channels per given time.

A separate image was created for each of the individual channels by reshaping the vector into a 256 pixels × 256 pixels image. The IR images were then subsequently circularly shifted to match the FOV of the visible image.

The general scheme was to segment the FOV utilizing the image of high resolution (the visible image), and use the stack of IR images to classify the chemical composition of each of the 256 × 256 pixels. Different classification methods were analyzed: K-means clustering and artificial neural network (ANN) classification was performed in Matlab, all the other machine learning methods were conducted employing sklearn [10] python packages. For some of the samples, segmentation was performed first, and then classification was pooled from the group of pixels belonging to the same segment, thus increasing SNR. The segmentation method determined for the respective samples is dependent on the morphological properties of the samples: for spherical samples such as beads, a circular shape was assumed. Therefore, for the PMMA and silica samples, a Matlab script was used to find circles first then the center and the radius of the circles within the FOV were calculated. For samples with irregular shapes, the watershed algorithm in OpenCV [11] was utilized to establish the different domains first.

5.3 Results and Discussion

5.3.1 Aligning Visible Beam and the IR beam

To compensate for the FOV offset caused by misalignment and systematic instrumentation drifts over time, post-processing of aligning the visible FOV and IR FOV was performed regularly. To do so, the two FOVs were superimposed and inspected when imaging a USAF (U.S. Air Force) test grid, and a horizontal and a vertical shift by the pixels was determined. Figure 5.2 illustrates how these shifts are reflected in realistic images. For this specific dataset, The IR images were circularly shifted 20 pixels horizontally and 25 pixels vertically, as determined by the USAF test grid. It is worth noting that the IR image has a considerably lower resolution because IR imaging was diffraction-limited. Therefore there is no optimal number of pixels to shift circularly.

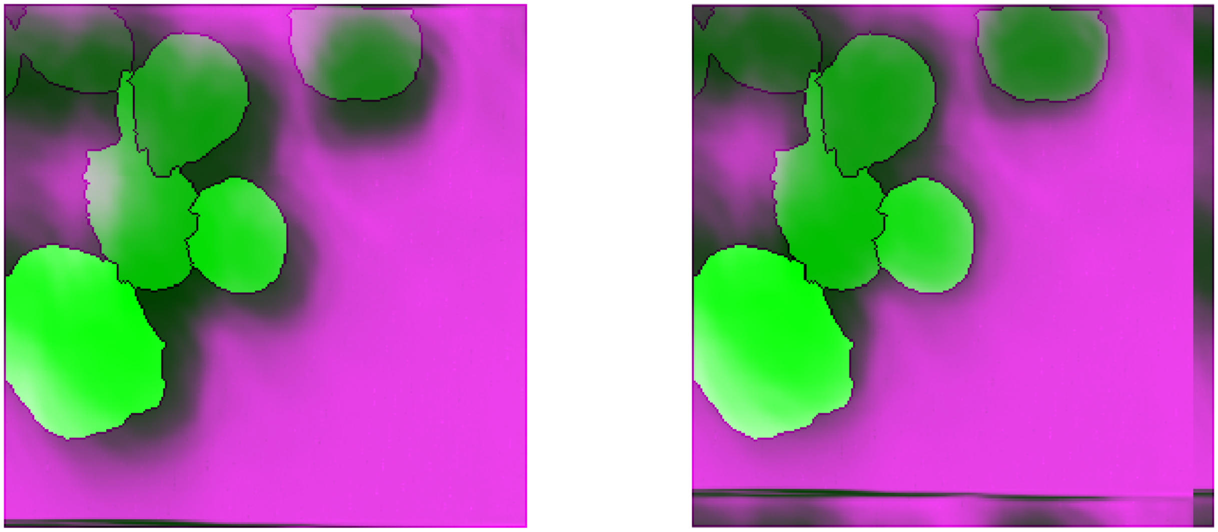
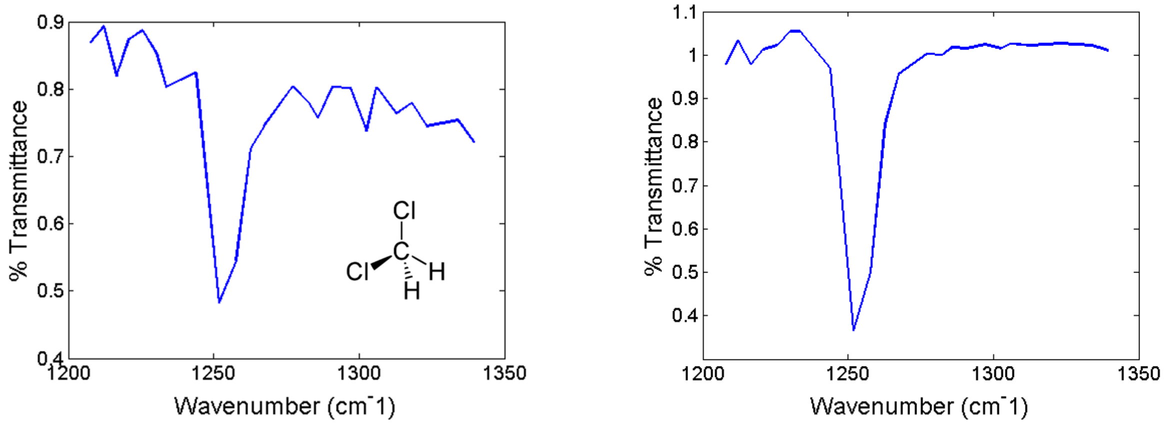


Figure 5.2. Before and after circularly shifting the IR images to overlay with the visible image

5.3.2 Spectroscopic Analysis of Dichloromethane

First, we confirmed the IR microscope's capability of acquiring an effective IR absorbance spectrum. The QCL channels have a wavelength range between 1190 nm^{-1} and 1340 nm^{-1} residing within the fingerprint region, making this technique holding potential for identifying numerous chemicals. Of the many IR absorbances within this region, the C-H oscillation exhibits high absorbance. The acquired spectra of DCM results are presented in Figure 5.3a (spectra of a single-cycle) and Figure 5.3b (averaged spectra from 50 cycles). A single spectrum was acquired from this microscope in as few as 9.6 µs, taking advantage of the QCL array firing 32 channels with a pulse width of down to 300 ns. The acquisition time constraint for the 50 times averaged spectrum was the duty cycle of QCL cannot be above 2% per channel, 20% overall. Accordingly it was configured to fire 200 ns pulses followed by 800 ns before the next channel. Thus the overall firing frequency was 1 MHz, resulting in a 1.6 ms acquisition time for 50 cycles.



(a)

(b)

Figure 5.3. Absorbance spectra for DCM. (a) Absorbance spectra for DCM acquired with a single-cycle. (b) DCM spectra averaged with 50 cycles.

As shown in Figure 5.3 both spectra captured the absorption at 1255 cm^{-1}, albeit the single-cycle spectra displayed a large noise ratio, (the noisy areas in the non-peak area), but the high noise level disappeared after taking an average of 50 cycles.

5.3.3 Segmentation Based on Sample Morphology

For simplicity reasons, images of beads of spherical shape were first segmented assuming perfect spherical shape, i.e., per segment can be described using a single parameter r (radius). Even though the actual beads may not be perfectly spherical, the deviation from the spherical shape is negligible. To segment this kind of sample, we first generate a dynamic threshold map using the function *adaptthresh*; the the image was then binarized and filled with holes using *imbinarize* and *imfill* respectively; finally, all of the circular shapes within the field of view were identified using *imfindcircles*. The segmentation effect of this method can be viewed in Figure 5.8. The key steps and functions are listed below:

```
T = adaptthresh(img_red,sensitivity);
img_bw = imbinarize(img_red,T*scale_factor);
img_bw = imfill(img_bw,'holes');
[centers,radii] = imfindcircles(img_bw,radiusRange);
```

For samples with more complicated morphology, such as the clopidogrel bisulfate crystals, the segmentation task becomes a more complex one. There are numerous sophisticated methods to create an optimal mask, many of which are complicated algorithms involving separate training and validation processes. Dilpreet *et al.* listed various image segmentation methods to be categorized into seven divisions: threshold-based, edge-based, region-based, clustering-based, watershed-based, PDE-based and ANN-based[12]. The recent research focus has been on ANN, with various applications in biomedical imaging. The multi-functional ANN algorithms can solve problems that call for separating similar objects that are contacting each other. However, ANNs are multi-processed and tedious to train. The watershed method was adopted and tailored for the clopidogrel samples for its generalizability and simplicity. It also offers several parameters for optimization, while no separate training process was needed.

The watershed algorithm considers the grayscale image as a topographic surface. Briefly, it establishes the low-intensity pixel areas as "valleys", and fills the isolated "valleys" with "water".[13] The watershed algorithm is a multi-step process, as shown in Figure 5.4. Prior to feeding the images into the algorithm, the images were inverted as signals collected by PMT had inverted intensity. To create a boundary between two contacting objects, figure 5.4a was hard-thresholded to generate a binary image, then taken distance transform (Figure 5.4b) – foreground pixels are replaced with distance calculations of the nearest background pixels – therefore locating the object centers of the foreground. Then these centers are inverted as "valleys" to be passed on to the watershed algorithm to "fill" – i.e. detecting the boundaries of the contacting objects (Figure 5.4c). The watershed approach generally works irregular domains, specifically for the clopidogrel bisulfate poly-crystalline; the particle domains are similar to spheres, with frequent cases of particles contacting one another.

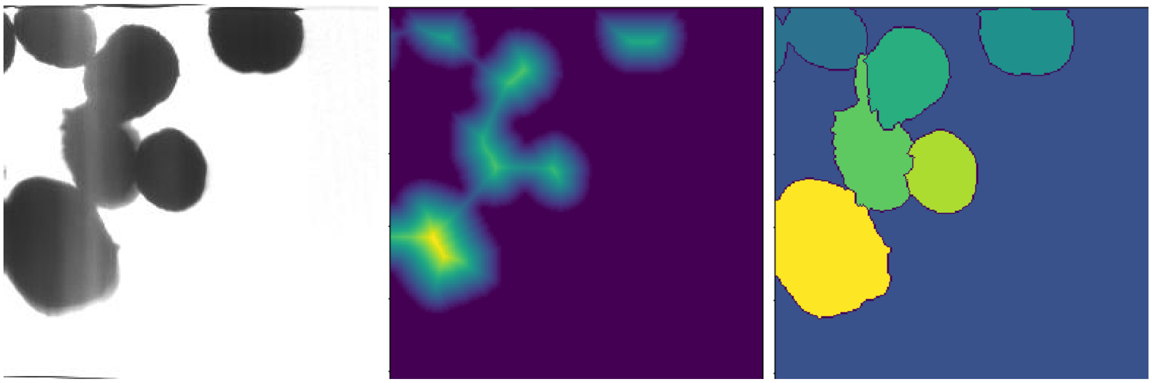

(a) Raw image, after inverting and scaling the intensity.
(b) Distance transform of Figure 5.4a.
(c) Final segmentation results

Figure 5.4. Segmentation of clopidogrel sample using opencv

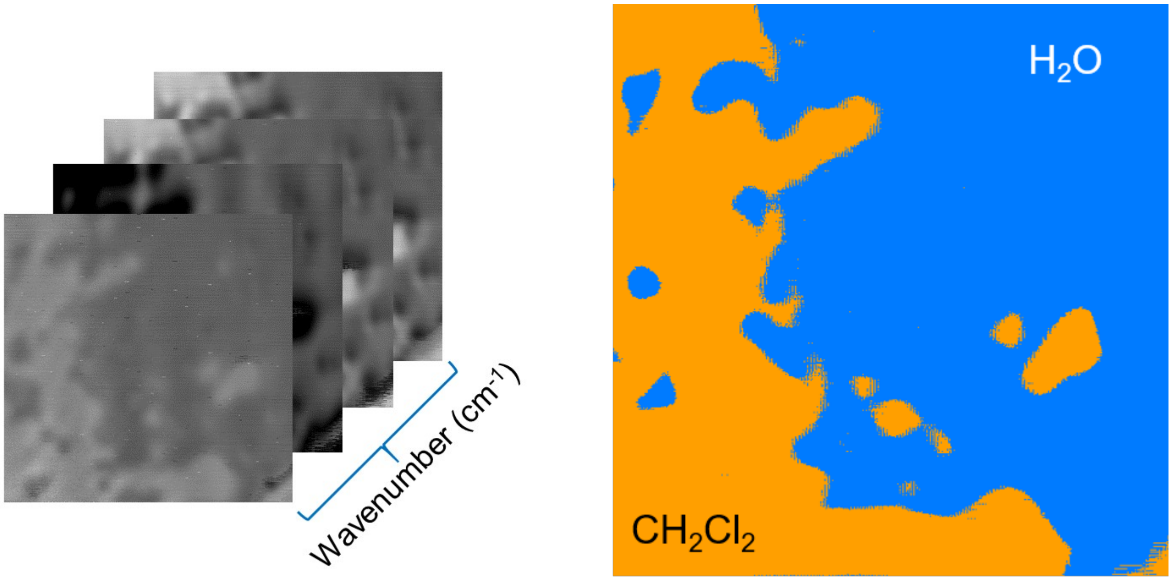
To obtain the finest segmentation results, looking at our samples, the types of parameters that can be fine-tuned are the cutoff values for thresholding to generate the initial mask, types of distance transform (L1, L2, etc.). For the majority of the clopidogrel samples, the default parameters offered respective functions mentioned previously in the dissertation, making this process an automatic fast approach.

Alternatively, a superpixel approach can be adopted. Briefly, the superpixel approach groups pixels similar in intensities[14], In other words, the images would be split into a user-defined number of discretization. However, the disadvantage of this method, is that the segmented image does not immediately resemble prominent shapes in the original image.

Therefore, the superpixel approach is not suitable for the analysis reported in this chapter. However, if the samples in question were not well-shaped crystals, but rather domains without clear boundaries, the superpixel approach might be an approach to utilize.

5.3.4 Classification of the Segmented Images

It has been demonstrated that DCM presents a clear spectral feature at 1255 cm^{-1} due to the C-H scissors mode (see Figure 5.3). Towards the goal of high-resolution IR images, for proof of concept, a stack of hyper-images was acquired as shown in Figure 5.5a. After applying K-means clustering, the interface between DCM and water was determined as illustrated in Figure 5.5b.



(a) (b)

Figure 5.5. Classification of DCM using k-Means. (a) When imaging, simultaneously acquired images at each wavelength from the array. (b) K-means clustering to distinguish DCM from H2O

Next, beads of different compositions (PMMA and silica) were imaged. Classifying based on IR transmittance of this sample is a more complicated case than DCM, which consisted of only even film surface. The IR transmittance will consist of partially scattered light because of the curvature of the beads. Nevertheless, the transmitted IR should still be specific to respective components.

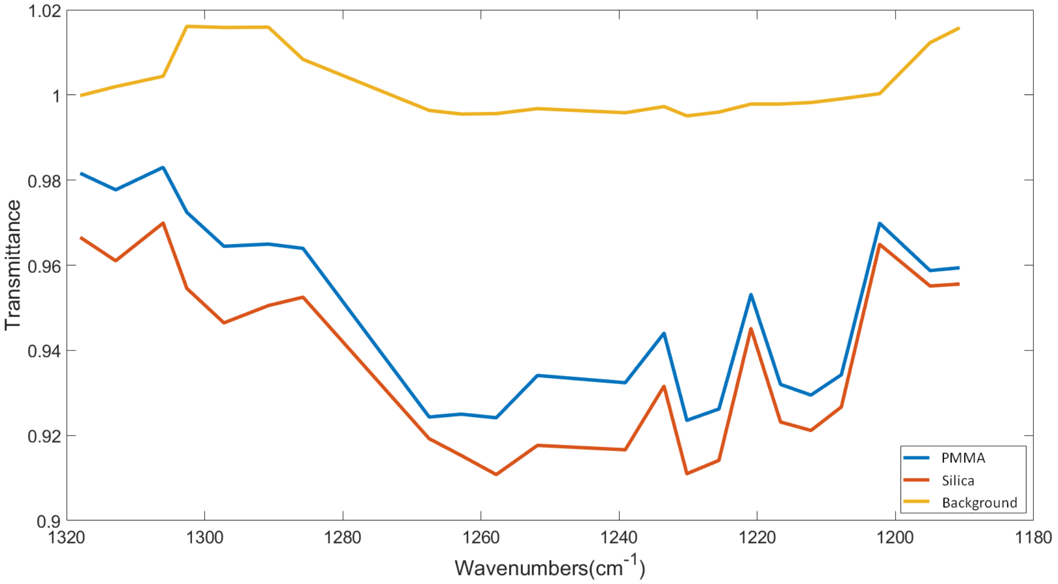


Figure 5.6. Spectrum of the beads

First, the beads were imaged separately with pure components – the images indicated that the averaged PMMA and silica spectra resemble each other in most spectral features (Figure 5.6). In addition, the small number of channels (32) limited the resolution of the spectra, making it harder to classify. However, the inherent distribution of the spectrum was not to be fully represented by the averaged spectra as shown in Figure 5.6, but rather, the inherent distribution of the total that the ANN would still capture. Therefore, a simple ANN was trained from the GUI of MATLAB for this classification task. This ANN contains just one hidden layer of ten neurons. Around ten particles of each class (PMMA, silica, or background) were cropped and reshaped into matrices as *Input*, with rows representing features and columns representing sample number. The ANN reported the receiver operating characteristic (ROC) plot as shown in Figure 5.7. ROC plot is a way of determining the quality of the ANN classifier (or any classifier). The more each curve hugs the left and top edges of the plot, the better the classification. According to Figure 5.7, the ANN was able to discriminate between the classes PMMA, silica, and background.

Using the trained ANN, a mixture image of PMMA and silica was fed into the ANN for a pixel-based classification, the results are shown in Figure 5.8. In this figure, 5.8a represents the visible image, and 5.8b one of the IR images. The visible image clearly displays the feature of the spheres, The IR image is diffraction-limited, which only displays blocks of shades, instead of clear features of either the silica or the PMMA. Note that the two lines (one horizontal and one vertical close to the edges) means that the IR images had been circularly shifted to match the FOV of the visible image.

The IR stack of images was then fed into the trained neural network, which in turn performed pixel-based classification, as shown in Figure 5.8c. The colors are manually assigned to correspond to different classes. Green is designated to represent glass, red to PMMA, and blue to background. Despite some individual pixels being mis-classified within the two spheres (green and red), the consensus is correct as corroborated by particle size. When merging the information obtained by the visible and the IR images, the visible image was segmented assuming spherical particles (see Section 5.3.3), then the consensus class from each of the individual segments was pooled, the final result of the classification as shown in Figure 5.8d.

Since the preliminary results with glass and PMMA beads are promising, the more interesting clopidogrel bisulfate sample was analyzed. Clopidogrel bisulfate is regularly used in conjunction with aspirin for various cardiovascular disease treatment[15]. Clopidogrel bisulfate has two polymorphs, Form I and Form II, corresponding two different molecular structures, with one of the forms (Form I) being the active drug ingredient. The same imaging process for silica and PMMA samples was repeated for clopidogrel mixtures, results are shown in Figure 5.9.

The clopidogrel sample is majorly spherical, yet as an actual crystal, there are not strictly spherical. Therefore, the previous segmentation methods based on that assumption do not hold, even though they worked for the beads. This segmentation task is a more complicated one. The detailed process to ideally segment this image without excessive assumption of trivial details, the detailed process is described in Figure 5.3.3. Note that, even with

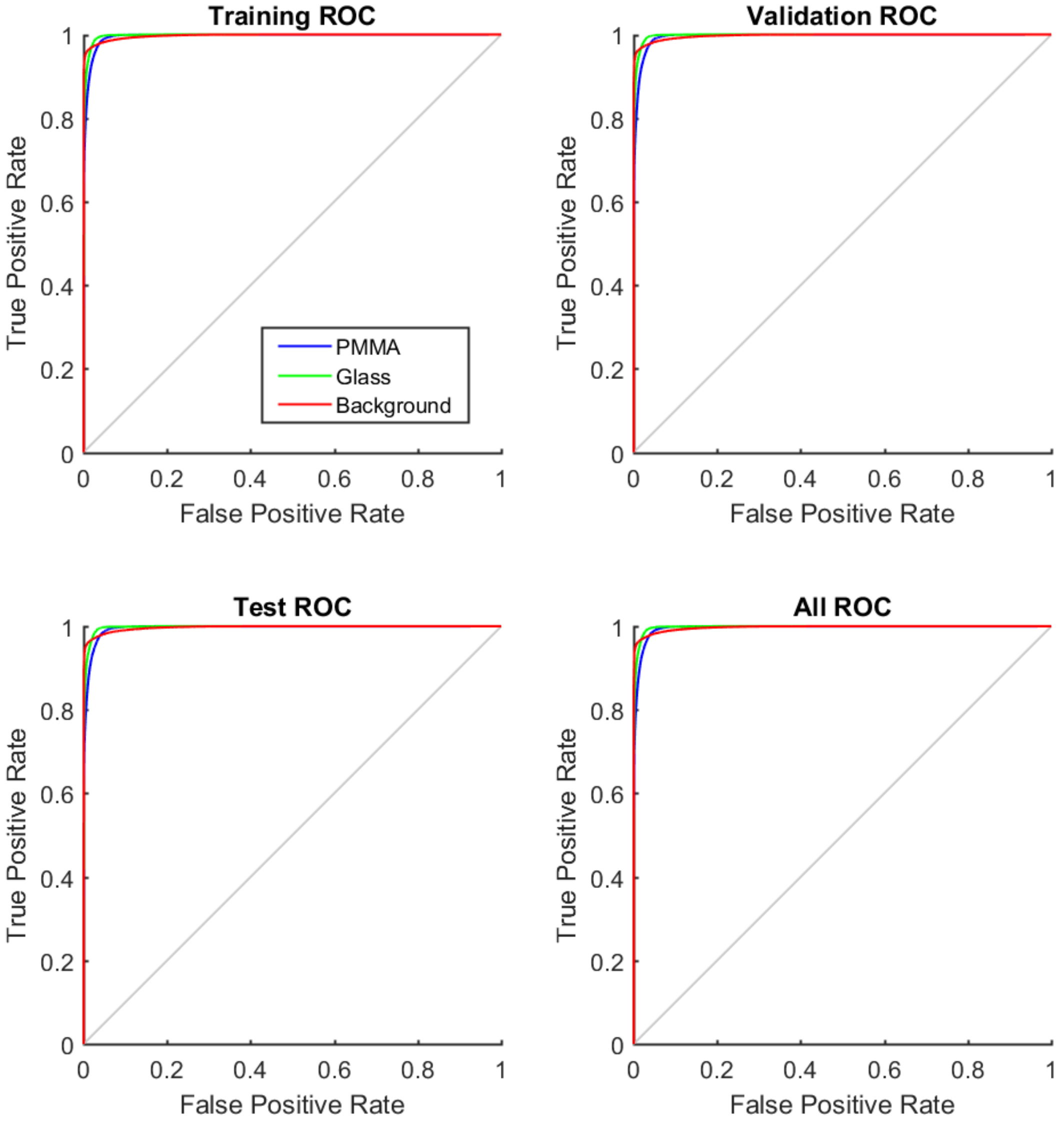


Figure 5.7. ROC plot reported by ANN

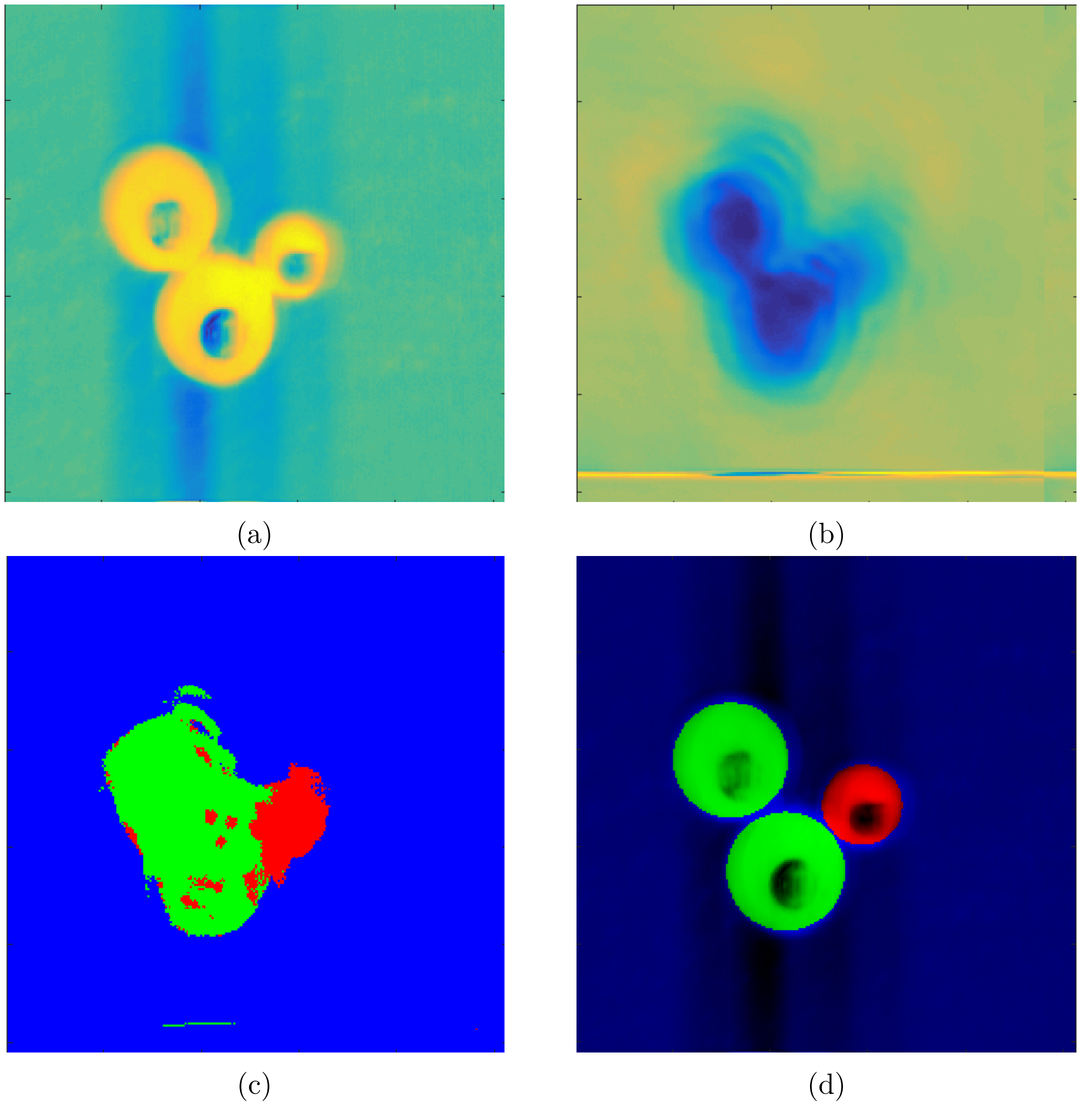


Figure 5.8. Images acquired via hyperspectral microscopy, unprocessed and classification results. (a) Visible image at 640 nm. (b) IR image at 8038 nm. (c) Pixel based ANN classification. (d) Processed classification result.

fine-tuning, the segmented image still wrongly labeled some impurity as part of the particle (bottom center particle). After fine-tuning the watershed algorithm, the result of segmentation is presented in Figure 5.9c. Pooling the classification results from a classifier produced the final result in Figure 5.9c, in which the original visible image is overlayed with false color as classification designation.

5.3.5 Classifiers Comparison

Apart from the complication induced by the non-trivial morphology of clopidogrel crystals, the decision of which classifier to use is equally, if not more complicated. Theoretically, the classifier can be any classifier; even non-supervised methods such as K-means can sometimes yield decent results, as is the case with DCM. However, the complication lies in the fact that the QCL laser beam, albeit portable, fast and tunable, has a limited wavelength range and limited channels, thus, limited spectral resolution. Therefore, the spectra acquired by the QCL microscope were of non-trivial distribution that would be challenging for the classifier to capture. Figure 5.10 is multiple selective visible images (arranged in a matrix), along with pixel based classification results produced by various off-the-shelf classifiers using scikit[10]. The types of classifiers demonstrated here are k-nearest neighbors[16], linear SVM[17], Gaussian process[18], decision tree[19], and random forest[20],

Classification performance is highly dependent on the characteristics of the input data. There are numerous source guides to which one to use depending on the property and quantity of the data[21], and new algorithms are constantly being developed (even though just as variations of existing algorithms), Furthermore, most guidances are empirically based[22]. For the hyperspectral dataset produced by the QCL microscope, the problem is unique in that the spectral data is not high dimensional per se, with the number of active channels being 22 in the clopidogrel bisulfate dataset, which is fewer than general spectroscopic data ranging from hundreds to thousands of features (even the majority of those features are redundant, this topic warrants another separate discussion) Therefore, it is useful to do an exhaustive search of the common algorithms. Figure 5.10 displayed just a subset of dozens of algorithms available off-the-shelf.

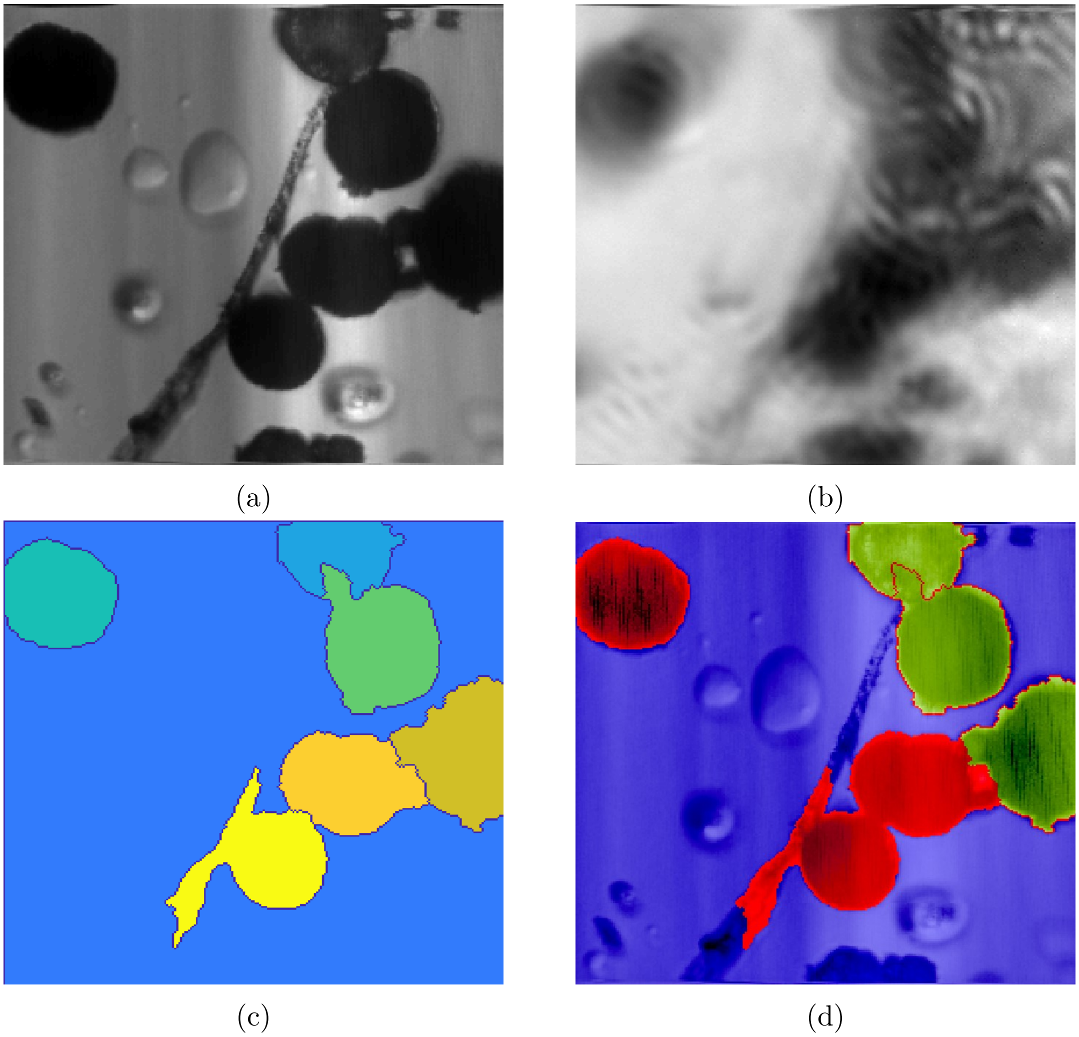
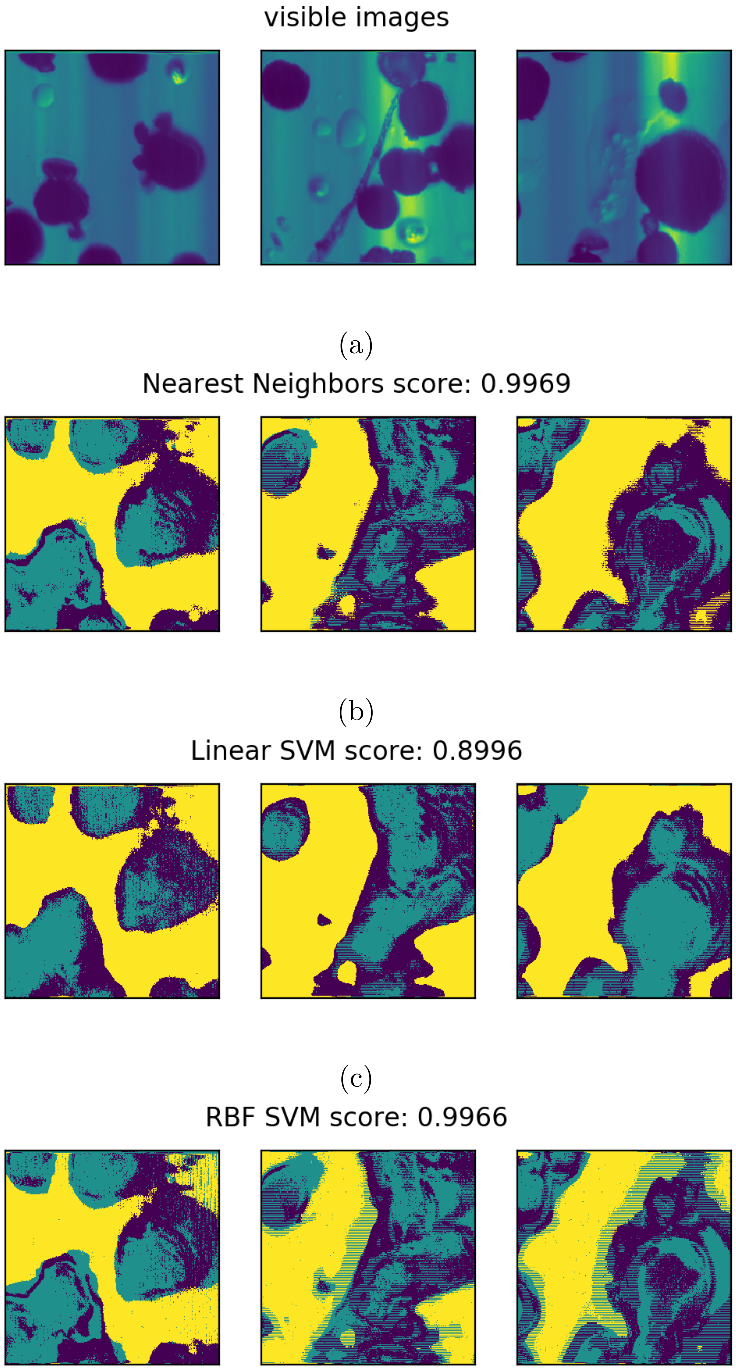


Figure 5.9. clopidogrel results (a) Visible image at 634 nm. (b) IR image at 8038 nm. (c) Segmented results based on Figure 5.9a. (d) Processed classification result.



visible images

(a)

Nearest Neighbors score: 0.9969

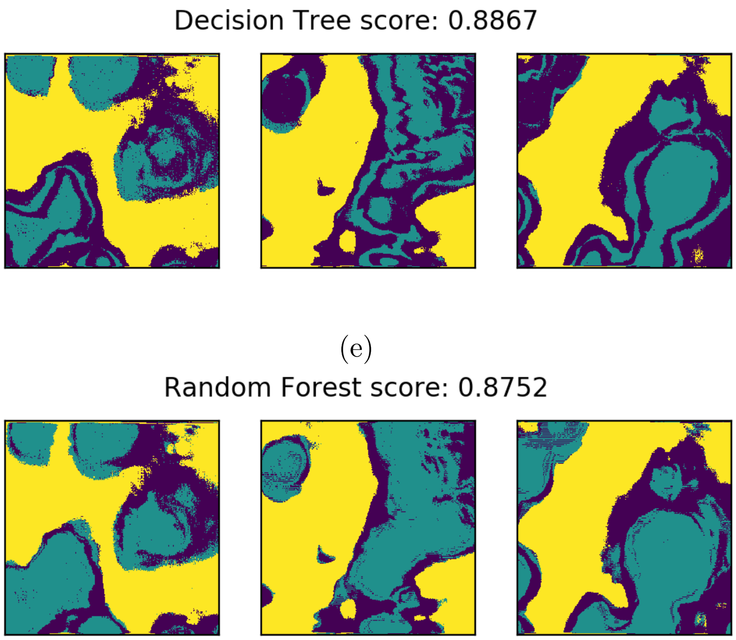
(b)

Linear SVM score: 0.8996

(c)

RBF SVM score: 0.9966

(d)



(e)

(f)

Figure 5.10. Visible images (a) and the pixel based classification results using various classifiers. (b) K-nearest neighbors, (c) linear SVM, (d) radial basis function kernel SVM, (e) decision tree, (f) random forest.

Upon close examination, all the algorithms agree on the classifications between the background and the sample, but disagree on the class assignment between the two polymorphs. For example, Figure 5.10e displays grooves of pattern that suggests the the classification was more intensity-based rather than spectral feature-based; Figure 5.10d seems to favor one class over the other (teal versus indigo), suggesting systematic error within the algorithm Each classifier reports a score (depending on specific classifier's definition of a "score", mostly score will be *accuracy*, number of correctly predicted data points out of all the data points) of between 0.85 to 0.99, meaning the classifier is confident of its prediction, yet the final results vary hugely, making it a tough choice for selecting one algorithm over the other. Generally speaking, the diagnostics for choosing an estimator is time-consuming. The practice often is to build a quick-and-dirty prototype; then, the focus is shifted to diagnose and fix the problem based on error analyses and ablative analyses. For the case with this clopidogrel bisulfate, the decision tree was the classifier that produced the results in Figure 5.9(d).

Alternatively, for future development in the research area of hyperspectral classification for IR images, a deep neural network can be constructed, as shown in the previous chapter (Section 4.2.2), the number of hidden layers, number of neurons per layer, choice of activation function and the remaining of the hyper-parameters can be optimized via Tensorboard[23] iterative testing, in which performance can be visualized with regard to each set of hyper-parameters.

5.4 Conclusion

A hyperspectral microscope was constructed to provide multiple imaging modalities. 32 IR channels with a spectral range of 1190–1340 cm^{-1} produce 256 pixel × 256 pixel × 32 channel hyperspectral image stack. Initial efforts demonstrated spectroscopy capable of distinguishing DCM and water, as well as a corresponding hyperspectral image. Image intensity encoded by the laser transmittance image fused results from visible and infrared images. Segmentation methods regarding sample morphology were discussed in detail and different recommendations were made based on the choice of sample. Merging IR images with 634 nm bright-field images fused spectral information with visible-wavelength resolution. Different classifiers were evaluated for this spectral dataset and the choice of the classifier was discussed in detail.

References

[1] J. K. Guillory and R. I. Poust, "Chemical kinetics and drug stability," *Modern pharmaceutics*, vol. 4, p. 142, 2002.

[2] D. E. Bugay, "Characterization of the solid-state: Spectroscopic techniques," *Advanced Drug Delivery Reviews*, vol. 48, no. 1, pp. 43–65, 2001.

[3] R. Mendelsohn, C. R. Flach, and D. J. Moore, "Determination of molecular conformation and permeation in skin via ir spectroscopy, microscopy, and imaging," *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1758, no. 7, pp. 923–933, 2006.

[4] S. A. Tofail, A. Mani, J. Bauer, and C. Silien, “In situ, real-time infrared (ir) imaging for metrology in advanced manufacturing,” *Advanced Engineering Materials*, vol. 20, no. 6, p. 1 800 061, 2018.

[5] S. L. Morissette, Ö. Almarsson, M. L. Peterson, J. F. Remenar, M. J. Read, A. V. Lemmo, S. Ellis, M. J. Cima, and C. R. Gardner, “High-throughput crystallization: Polymorphs, salts, co-crystals and solvates of pharmaceutical solids,” *Advanced drug delivery reviews*, vol. 56, no. 3, pp. 275–300, 2004.

[6] R. Bhargava, “Infrared spectroscopic imaging: The next generation,” *Applied spectroscopy*, vol. 66, no. 10, pp. 1091–1120, 2012.

[7] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, “Quantum cascade laser,” *Science*, vol. 264, no. 5158, pp. 553–556, 1994.

[8] C. Kuepper, A. Kallenbach-Thieltges, H. Juette, A. Tannapfel, F. Großerueschkamp, and K. Gerwert, “Quantum cascade laser-based infrared microscopy for label-free and automated cancer classification in tissue sections,” *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[9] K. Yeh, S. Kenkel, J.-N. Liu, and R. Bhargava, “Fast infrared chemical imaging with a quantum cascade laser,” *Analytical chemistry*, vol. 87, no. 1, pp. 485–493, 2015.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in {P}ython,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, 2000.

[12] K. Dilpreet, K. Yadwinder, D. Kaur, and Y. Kaur, “Various Image Segmentation Techniques: A Review,” *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 5, pp. 809–814, 2014.

[13] L. Vincent and P. Soille, “Watersheds in digital spaces: An efficient algorithm based on immersion simulations,” *IEEE Computer Architecture Letters*, vol. 13, no. 06, pp. 583–598, 1991.

[14] D. Stutz, A. Hermans, and B. Leibe, “Superpixels: An Evaluation of the State-of-the-Art,” Tech. Rep. arXiv: 1612.01601v3.

[15] G. Patti, G. Micieli, C. Cimminiello, and L. Bolognese, “The Role of Clopidogrel in 2020: A Reappraisal,” *Cardiovascular Therapeutics*, vol. 2020, 2020, ISSN: 17555922. DOI: 10.1155/2020/8703627.

[16] J. M. Keller and J. A. Givens, “A fuzzy k-nearest neighbor algorithm,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-15, no. 4, pp. 580–585, 1985, Cited By :1567. [Online]. Available: www.scopus.com.

[17] J. Platt, *Probabilistic outputs for svms and comparisons to regularized likehood methods, advances in large margin classifiers*, 1999.

[18] C. E. Rasmussen, “Gaussian processes in machine learning,” in *Summer school on machine learning*, Springer, 2003, pp. 63–71.

[19] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, “Classification and regression trees. belmont, ca: Wadsworth,” *International Group*, vol. 432, pp. 151–166, 1984.

[20] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] D. Sarkar, R. Bali, and T. Sharma, “Practical machine learning with python,” *A Problem-Solvers Guide To Building Real-World Intelligent Systems. Berkely: Apress*, 2018.

[22] D. Chicco, “Ten quick tips for machine learning in computational biology,” *BioData mining*, vol. 10, no. 1, pp. 1–17, 2017.

[23] Martıén Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/.

VITA

Youlin Liu went to University of Science and Technology of China for undergraduate school, where she enrolled in biology, and switched to chemistry after the first year. Her first research project was in Dr. Hanqing Yu's lab on developing biological solutions for water contamination monitoring. She later joined Dr. Zhaoxiang Deng's lab; she optimized the synthesis of nano-structure for optical applications. In 2015, Youlin was founded by China Scholarship Council to conduct three-month research with Dr. David Fernandez Rivas in The University of Twente, the Netherlands. Under Dr. Rivas, she researched microfluidics using sonochemistry as a probe, on a micro-reactor that later went commercialized. In 2016, Youlin attended Purdue University and joined Dr. Garth Simpson. Her research projects vary from instrumentation of optical microscopes and algorithms& method development for analytical chemistry data incorporating machine learning. Youlin graduated with her Ph.D. in Analytical Chemistry from Purdue University in August 2021.