Research Summary

Youlin Liu

April 5, 2021

During my PhD research, I have been working with instrumentation building and the interpretation of spectroscopic data. Some of the projects are pursuing Machine Learning methodology as complementary to traditional chemometrics (GALDA); some of the projects are heavily hands-on electronics and optics (Hyper-spectral imaging); in collaborative projects my role was to build mathematical model (High-Throughput FRAP) and X-Ray data indexing (Synchrotron X-Ray Damage Analysis). The unifying theme is that I work to optimally interpret complex chemistry data.

1 Generative Adversarial Linear Discriminate Analysis

Algorithm 1: Generative Adversarial Linear Discriminate Analysis (GADLA) algorithm

Input: Excel Spreadsheet
Output: Optimized LDA loading
Data: Spectroscopic data set

```
/* Data preprocessing                                    */
data ⇐ read from excel
training set, validation set ⇐ data
/* on training set                                       */
spectral mean and standard deviation of each class
  ⇐ training set
perform n-class LDA
/* n = number of classes in data                         */
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
     4th class
    perform (n+1)-class LDA
    if LDA loading smoothed to desired extent then
        break while loop
    else
        keep iterating
/* validate LDA loadings with validation data            */
```

Modern instrumentation can produce ever-increasing volumes of measurements, the collective analysis of which is routinely used to inform decision-making. As the dimensionality of the data increases, the greater becomes our collective reliance on algorithmic data analysis approaches for quantification, classification, and basic interpretation. The common approaches for feature extraction and dimension reduction are with many caveats in implementation, some of which are prone to overfitting (neural networks), others, an effective result relies on rigorous cross validation (PLS-DA). In short, models often require careful tuning of parameters and are not best representative of sample population.

In light of the successes of generative adversarial approaches to minimize over-fitting artifacts in ANNs, we hypothesize that analogous benefits may be realized to address the numerical instabilities complicating the use of LDA at full rank 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 rank LDA to optimally separate genuine and generated training data, iteratively optimized to compete against each other. The algorithm structure is shown in Algorithm 1.

The performance of GALDA was compared with other common dimension reduction methods of spectral data, the results of which are summarized in Figure 1. Specifically, alternative methods included PCA, PLS-DA, PCA+LDA, and regularized LDA using a shrunken centroids regularizer. Two key metrics were considered for assessing the merits of the different dimension reduction approaches: resolution and overfitting. 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.

In conclusion, GALDA yields good resolution, compa-

rable to standard methods, while maintaining low overfitting level, regardless of statistical distribution of the data.

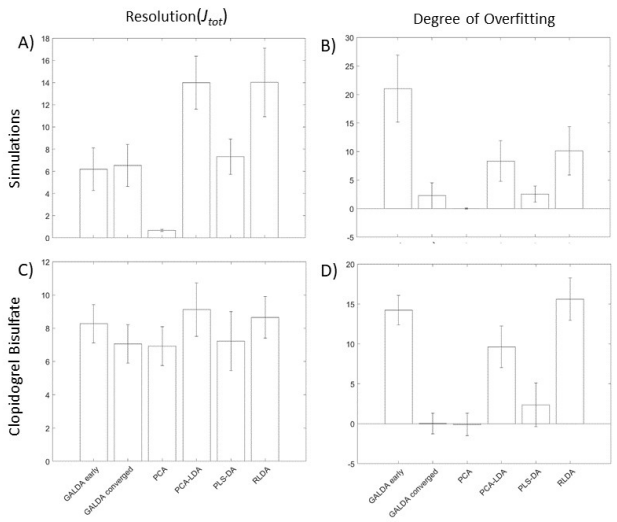

Figure 1: Cross-validated comparison results of GALDA with other standard methods

2 Hyperspectral Infrared Imaging Microscope Design and Digital Image Analysis

In this work, a hyperspectral microscope operating in the LWIR is demonstrated for future application to stand-off imaging platforms. It is our aim to integrate sparse-sampling methods, into our imaging system to improve the accessible frame rate. We hope to extend our current effort in microscopy to stand-off imaging to improve the frame rate of IR imaging for remote threat detection.

A QCL array was chosen as the light source for this system. QCL sources emit light at IR wavelengths, operate at room temperature, and produce milliwatts of radiation. QCL sources are also lightweight and have low power requirements, making them ideal for systems that are being developed for security applications. The QCL microscope follows the design as shown in Figure 2.

Figure 3 shows the imaging results of this microscope. The co-propagating visible and IR beam acquire transmittance images separately, of the same field of view. The image acquired by the 640 nm laser was first segmented, each of the segmented area were pooled for classification using the corresponding IR pixels. The final result in (D) is the combined information of spatial and IR image.

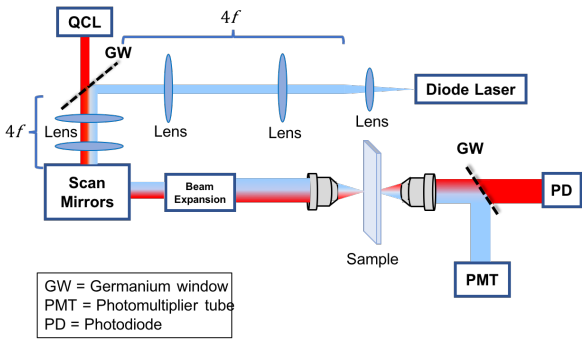

Figure 2: Instrument schematic of the LWIR hyperspectral microscope. 640 nm beam is emitted from the diode laser and is combined with 32-channel LWIR beam emitted from the QCL.

3 High-Throughput Fluorescence Recovery After Photobleaching Diffusion Analysis of Protein/Excipient Interactions

Therapeutic macromolecules including monoclonal antibody (mAb) drug conjugates represent approximately half the total global market for new chemical entities. Subcutaneous (SC) delivery of therapeutic macromolecules when compared to the established intravenous (IV) route, has many issues regarding incomplete bioavailability. As such, diffusion measurements in well-characterized matrices highly representative of those anticipated within subcutaneous environments may provide early-stage in vitro assessments of anticipated *in vivo* bioavailability for subcutaneous injection of potential therapeutic mAb candidates.

In this proof-of-concept work, we combine robotic low-volume (700 nL) sample preparation with high-throughput fluorescence recovery after photobleaching (FRAP) measurements to evaluate labeled protein mobility within model matrices prevalent within the subcutaneous environment. We also extended the DLVO theory that was initial introduced to describe protein-protein interactions to integrate protein/matrix interactions, allowing interpretation of the measured trends

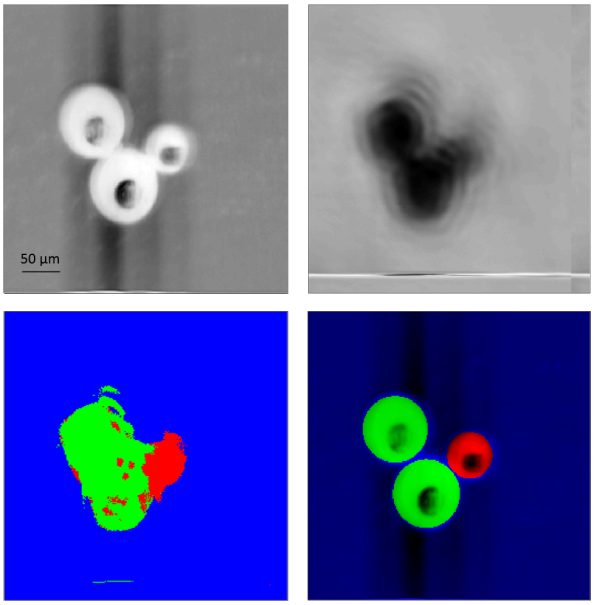

Figure 3: Images acquired via hyperspectral microscopy are displayed for the red laser transmittance (A), a single QCL channel (B), the classified IR image (C), and the segmented classified hyperspectral image (D). Enhanced spatial resolution of spectral information is obtained by combination of visible and IR wavelengths.

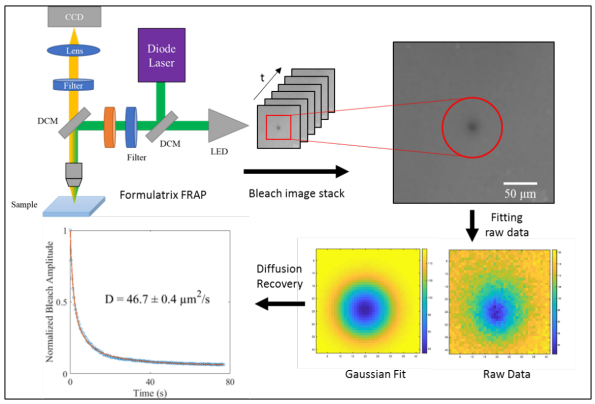

Figure 4: Schematic illustration of high-thoughput FRAP instrumentation and data analysis.

in diffusion in terms of long-range electrostatic interactions and short-range affinity. Independent measurements of size and charge for the proteins and biopolymers of interest allow diverse panels of diffusion measurements to be combined through least-squares fitting to recover a short-range affinity parameter proportional to the ratio of two contributions to the second virial cross-coefficients (excluded volume contribution divided by the contribution from short-range interactions). The protein-specific and matrix-specific parameters recovered experimentally allow prediction of protein mobility within matrices containing collagen and hyaluronic acid as a function of protein concentration, matrix concentration, and pH. Experimental design is shown in Figure 4.

Figure 5 contains the pooled results from all queried conditions from all 5 proteins investigated. These combined results suggest a clear correlation between the measured diffusion coefficient by FRAP and the predicted coefficient from DLVO theory evaluated with a single-parameter fit to the short-range adhesion parameter. Despite the simplicity with which short-range interactions were modeled and the fairly narrow range in diffusion coefficients, the trend correlating predicted and measured diffusion coefficients by DLVO is clearly

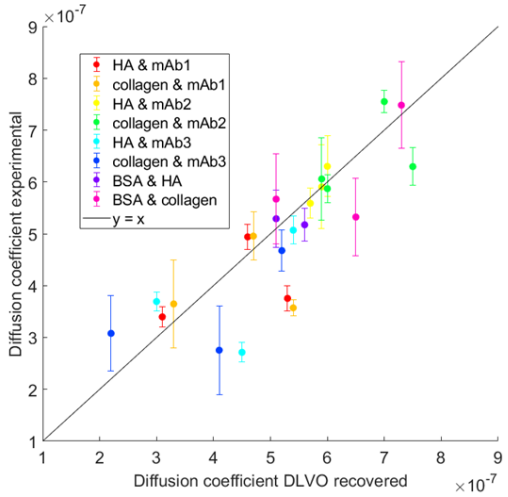

Figure 5: DLVO-recovered diffusion coefficient against the experimental ground truth diffusion constant.

evident in the figure.

4 Synchrotron X-Ray Damage Analysis With Non-Negative Matrix Factorization

Currently, X-ray diffraction (XRD) serves as the most widely used approach for the generation of high-resolution structures. However, damage induced by X-ray radiation ultimately limits the achievable resolution, by altering the sample during the measurement. These effects are exacerbated by advances in ever brighter X-ray sources enabling analysis of ever smaller crystals.

A substantial body of work has been directed toward mitigating radiation damage induced in protein crystals during XRD. As crystal sizes are reduced, current approaches and methods will be increasingly challenged. As such, computational methods designed to identify and remove X-ray-induced artifacts in recovered scattering patterns nicely complement experimental methods such as cryo-protection and X-ray beam-shaping.

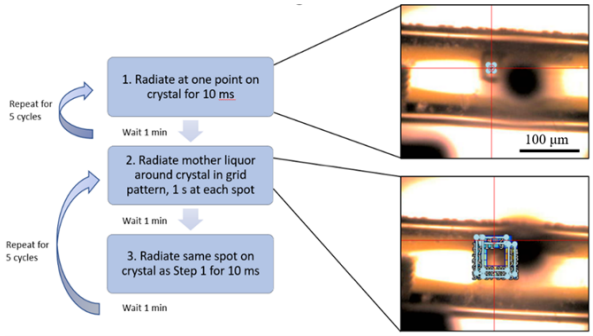

Figure 6: Schematic of the method used for data acquisition of room temperature XRD of lysozyme crystals, with images showing the area selected for radiation.

Here we describe a computational method for numerically isolating unperturbed peaks and monitoring the progression of chemically specific damage introduced upon X-ray exposure using non-negative matrix factorization (NMF). NMF enabled decomposition of dose-dependent scattering patterns into a series of components based on the physical requirements of non-negative intensities for the individual reflections and non-negative amplitudes of the components. This analysis provided additional insight into the mechanism of radiation damage and enabled isolation of the initial unperturbed XRD pattern. The experimental data collection is detailed in Figure 6.

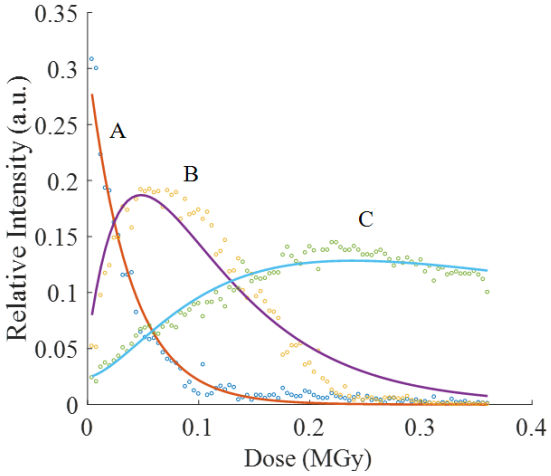

Figure 7: Kinetics of propagation of damage with X-ray dose in tetragonal lysozyme.

Upon NMF , the kinetics for dose-dependent evolution of the different components in the crystal is shown in Figure 7. No kinetics constraints were imposed on the raw data when using NMF to select the different components. Nevertheless, the emerging sequential progression from A $\rightarrow$ B $\rightarrow$ C $\rightarrow$ D (nondiffracting) that is clearly evident from cursory inspection of the dose-dependent behavior is striking. These trends strongly suggest a sequential series of structural changes induced upon X-ray absorption.