Identifying Chromophore Binding Modes through Principle Component Analysis of FTIR Spectroscopy

Morgan Zemaitis, Taylor Moot, Rohan Isaac, Shannon McCullough, Rene Lopez, James Cahoon

Senior Honors Thesis

Curriculum for Environment and Ecology and the Department of Chemistry

The University of North Carolina at Chapel Hill

April 10, 2017

Approved:

Jim Cahoon, Thesis Advisor

Matthew Lockett, Reader

Abstract

In the quest for solar technology to reach widespread implementation, dye-sensitized solar cells (DSSC) have been studied as a cost effective alternative to silicon solar cells.[1] Here we use lead titanate ($PbTiO_3$), a novel p-type semiconductor, to probe the surface interaction of the P1 chromophore (4-(bis-{4- [5-(2,2-dicyano-vinyl)thiophene-2-yl]phenyl}amino)benzoic acid) through its binding modes, which can impact the charge transfer efficiency of the device. The identity of binding modes are typically found through a combination of FTIR, XPS, and Raman spectroscopy—all of which are predicated on high chromophore loading to give high signal to noise spectra. Yet p-type devices do not typically have as high dye loading as n-type and therefore lack research in this area. To facilitate understanding of these surface interactions, we use principal component analysis (PCA) of FTIR spectroscopy to identify binding modes at the P1-$PbTiO_3$ interface. This multivariate statistical method reduces large datasets to a lesser amount of principle components that separate out noise and capitalize on unique phenomena in the data without the prerequisite of prior assumptions. Although PCA used in this experiment proved to be informative, it renders incapable of identifying possible binding modes for P1.

1. Introduction

One of the largest challenges facing the energy industry is providing carbon-free electricity that has the reliability of traditional energy sources. A common criticism of renewables that are implemented on a large scale, specifically solar and wind technology, is their intermittency for megawatt production. Thus the scientific world has steered towards the direction of storage: how can this electricity be saved for times when the sun doesn't shine and the wind doesn't blow?

A dye-sensitized photoelectrosynthesis cell (DSPEC) can provide an alternative for solar energy storage by creating solar fuel. By integrating molecular light absorption and catalysis with metal oxide semiconductors (photocathodes and photoanodes) that are stacked in tandem, solar fuel can be created. One component of tandem dye-sensitized solar cells is an efficient p-type photocathode, which can reduce carbon dioxide to produce fuel precursors.[2]

Dye-sensitized solar cells (DSSCs) are used as a simplified model of the DPSEC to help focus on a single variable. Although this solar cell design has seen respectable power conversion efficiencies from n-type semiconductors (n-DSSCs) above 12%, p-type devices (p-DSSCs) show only 1/6 of the n-DSSC performance.[3] There is also significantly less research on p-type photocathodes in comparison to the more common TiO_2 photoanodes; less than fifty studies have been published up to 2011.[2] The lower efficiency and lack of scientific understanding renders strong incentives to study p-type materials.

In order to see improvements in p-DSSCs, semiconductors alternate from NiO, which is the prevailing semiconducting material to date, must be considered. Based on materials informatics analysis and experimental validation, $PbTiO_3$ has shown to be a promising new p-type semiconductor. However, there is little knowledge on the surface interactions of lead titanate in a p-DSSC.

The semiconductor-chromophore interface can significantly impact recombination and thus DSSC performance. This interface is the site where reduced dye molecules and holes in the semiconductor can recombine, which can impact the short circuit photocurrent density (J_{sc}), the open-circuit voltage (V_{oc}) and the fill factor (*ff*), all of which are measurements that indicate power conversion efficiency (η) of the solar cell.[2,4] More specifically, the bonding interactions and between the chromophore and the semiconductor can dictate the efficiency of charge transfer and reduced recombination. For example, maximizing the distance between the semiconductor surface and the lowest unoccupied molecular orbital (LUMO) of the chromophore can reduce recombination because of the distance between the photo-generated charges.[2]

A chromophore can bind to a semiconductor surface through a number of different binding modes, but research on this property has been primarily limited to n-type materials.[5] Research focused on p-DSSC properties have been hindered for a number of reasons, including by low dye loading[6,7] and the aforementioned general lack of extensive research on p-type materials.

In this experiment we focus on defining the interactions between the $PbTiO_3$ semiconductor and P1 chromophore. P1 has become a widely used organic donor-π-acceptor dye created for p-DSSCs. The electron donor is triphenylamine moiety and the electron acceptor is the malonitrile moiety, of which there are two.[8]

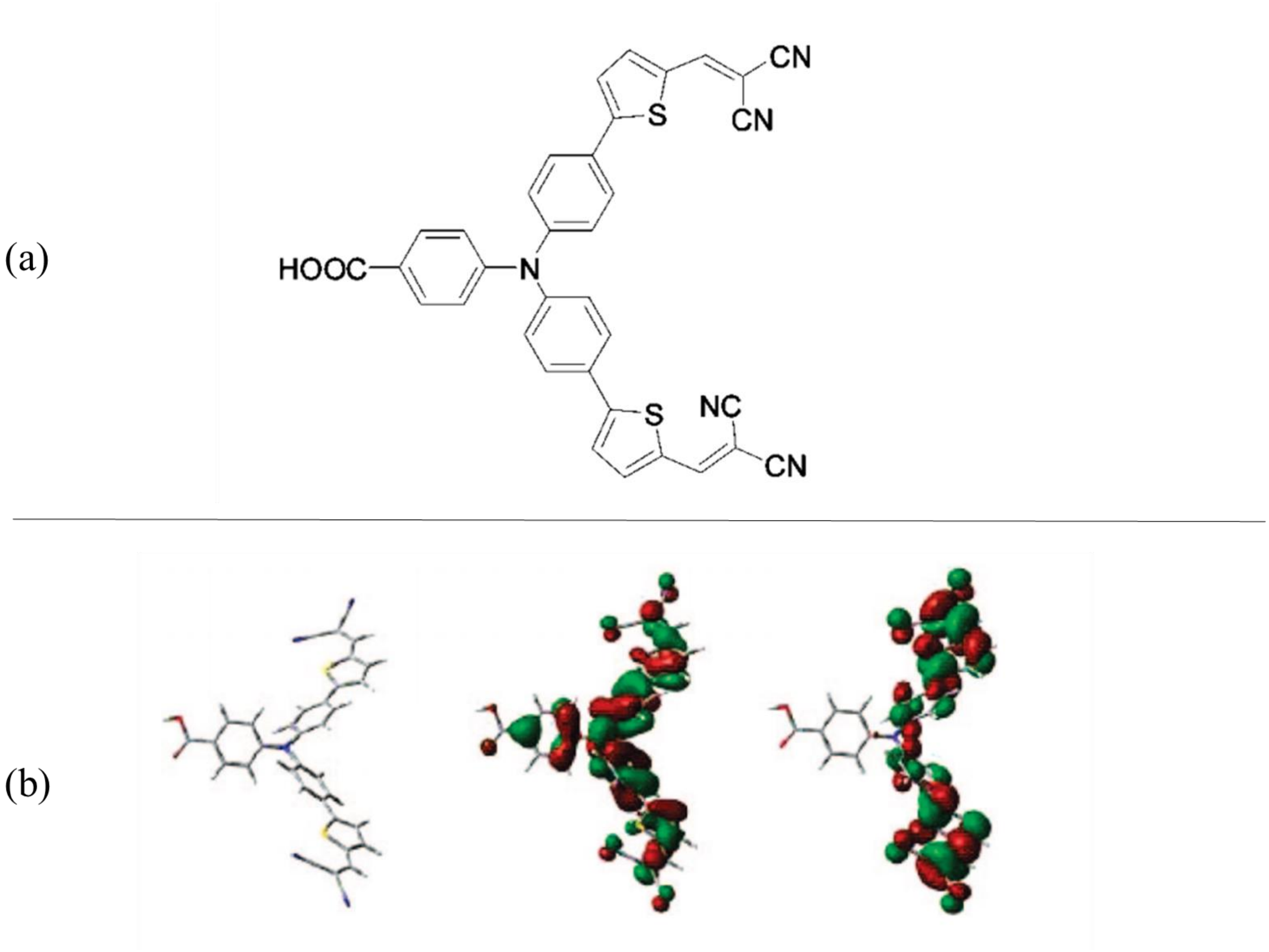


Figure 1: Part A shows the structure of the P1 molecule (4-(bis-{4- [5-(2,2-dicyano-vinyl)thiophene-2-yl]phenyl}amino)benzoic acid). Part B shows the optimized structure in the leftmost image, and the middle and right images show the frontier molecular orbitals of the HOMO and LUMO, respectively.[8]

There can be a multitude of chromophore binding modes at the P1-$PbTiO_3$ interface. It is well understood that most dyes in DSSCs use a carboxylic acid anchoring group that connects to the semiconductor surface.[9] but what is not known is exactly how this anchoring group binds to

$PbTiO_3$. In the case of TiO_2 and N719 dye molecules, the possible anchoring modes of the carboxylic acid identified through research include monodentate ester-type, bidenate chelating, and bidentate bridging.[5]These iterations can be true for the carboxylic acid anchoring group at the anchor-electrode interface for $PbTiO_3$ as well.

To understand the P1 binding modes on a $PbTiO_3$ surface, Fourier transform infrared (FTIR) spectroscopy, UV-Visible spectroscopy (UV-Vis), and density functional theory (DFT) calculations are used. Each of these techniques play a role in characterizing the binding modes, yet some of their limitations require more extensive data analysis.

FTIR is a commonly used method to measure the vibrational absorption spectra of molecules. Using a broadband blackbody radiation source, the spectrometer measures absorption from a wide wavelength range simultaneously.[10] For this study we also use attenuated total reflectance (ATR) which measures spectra through an internal reflection plate or prism instead of placing a sample directly into the IR beam.[11] Using this technique provides the advantage of preserving the films for other measurements such as UV-Vis or for direct use in DSSCs. The FTIR-ATR spectra obtained from experimentation characterize the vibrational structure of the molecules—its chemical components, groupings, and the types of bonds formed. The data and information obtained from the FTIR spectrometer is the cornerstone of binding mode identification for this experiment.

DFT is a type of computational calculation that uses quantum mechanics to approximate wave functions and equilibrium bond lengths and angles. The calculated spectra of P1 from this method is informative for indicating the ideal FTIR spectra which then determines the wavelength regions of interest in the experimental FTIR data. This is especially useful in cases where little is known about the experimental spectra of the molecule; to date there is no reported FTIR spectra of P1.

Where FTIR focuses on the infrared range of the light spectrum, UV-Vis measures the near ultraviolet, which ranges from 200-800 nm. Absorbance peaks in this spectrum indicate the visual color and electronic structure of the molecules as well as the aggregation of dye on the surface, based on any asymmetry or change in peak wavelength of the absorption peak. This information will support the FTIR data and DFT calculations by probing P1 aggregation.

In theory, the FTIR spectra informed by the DFT calculation and the UV-Vis spectra should give sufficient information of the P1-$PbTiO_3$ binding mechanisms. However, low dye loading on p-type surfaces causes poor signal to noise in the experimental data. Without the informative FTIR peaks of P1, the ability to identify its chromophore binding mode let alone its very presence is hindered.

Statistical analysis of spectroscopic data can identify trends and information even within datasets that contain minimal variance as a result of low dye loading. For this experiment, two data analytics methods were used: data preprocessing to remove errors and noise in the data and principal component analysis to identify important characteristics of the data variance.

The application of preprocessing techniques to spectra is well utilized – there is a diverse range of literature on preprocessing methods that pertain specifically to IR spectra.[12] There are a number of ways to address data errors and inconsistencies – some transform the entire datasets and others simply remove outliers. The workflow for preprocessing that was done for this experiment is indicated in Figure 2. After visualizing the raw data any spectra that seemed

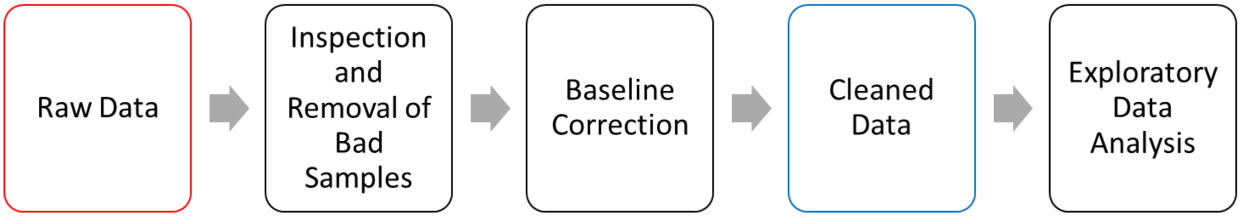

Raw Data
Inspection and Removal of Bad Samples
Baseline Correction
Cleaned Data
Exploratory Data Analysis

Figure 2: Data preprocessing workflow for experimental design.

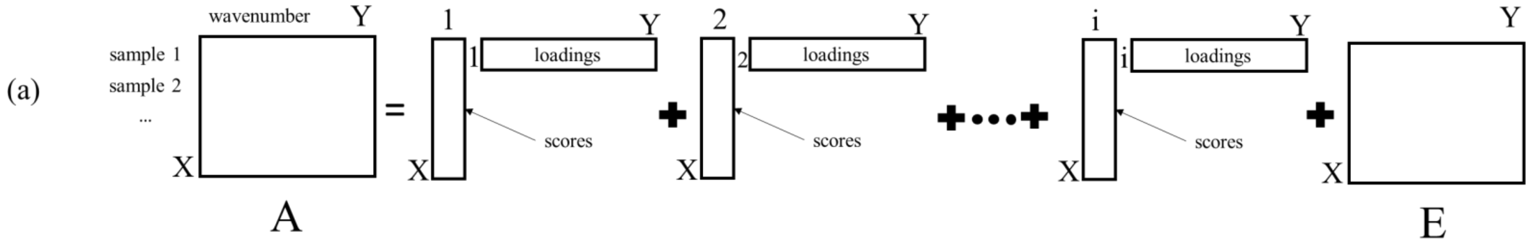
fundamentally different from the majority of the dataset, such as spectra with significantly different baselines, were removed.

Once the outliers are removed, the spectra are then examined for unstable baselines. It is not uncommon for experimental spectroscopy data to have this issue; its causes can come from atmospheric conditions, background measurements, or other factors. Baseline correction that compensates for these discrepancies usually consists of manually selecting points on the spectra as the beginning and end of peak regions which are then calculated by the computer. For the large amounts of spectra seen in this experiment, this method is time-consuming, subjective, and inconsistent. Instead, an automated function for baseline correction was used for each sample.

In the case of this dataset, asymmetric least squares (ALS) was used as the baseline estimator. This method was chosen based on the consistency and shape of the transformed spectra post-correction, which correlated with the DFT calculations more closely. ALS is an algorithm that uses a Whittaker smoother, which is a moving weighted average filter that removes high frequency data without the loss of information, to estimate the baseline and an asymmetric weighting of deviations from the predicted baseline to more accurately represent the dataset. By this calculation, the analytical peak signal stays intact while the noise and warping of the spectra is reduced.[13]

Once the data has been cleaned of noise and erroneous spectra, PCA can be used to conduct multivariate analysis. This method reduces large datasets into a lesser amount of principal components (PCs) which are used to distinguish unique phenomena in the variance between spectra. The information these PCs hold have the capability to classify objects that indicate similarities, often without prior assumptions of groupings or classification. These components are not correlated with one another, but they are weighted unequally: the importance of a PC variable in the model is dictated by the amount of variance it contains, and each component contains the maximum possible amount of variance in the data within the constraints of the model. This means that the first principle component represent a large majority of variance in the dataset, which each subsequent principle component (orthogonal to the preceding PC) representing the maximum amount of remaining variance.[14] The total number of principal components is equal to the amount of spectra being studied, i.e. if 10 samples were being studied then 10 principal components would represent 100% of the variance in data.[15]

For an experiment with spectroscopic data, a matrix **A** is created from **X** × **Y** where the **X** columns indicate the spectra for a specific wavelength of a singular sample and the **Y** rows represent each sample taken. Matrix **A** is then transformed into a scores matrix **S** and loadings matrix **L** and residual matrix **E**, as shown in Figure 3A. The scores and loadings matrices should contain the underlying patterns in the data and the residual matrix contains the noise.[16] Part B of Figure 3 shows this transformation linearly, where each term in the linear model represents a principal component that, summed with the residuals, should be equivalent to the original dataset. For the FTIR spectra used in this experiment, each loading vector l_i is the variance in spectra represented by the principal component and each score vector s_i represents the weight of the principal component spectrum for each of *i* samples.



(b)

$$A \quad = \quad s_1 l_1 \quad + \quad s_2 l_2 \quad + \quad \ldots \quad + \quad s_i l_i \quad + \quad E$$

Figure 3: Mathematical definition of PCA in Part A, and Part B is the matrix representation.[17]

Plots of the scores and loadings values describe visually the data variation that is represented in each principal component. Scores plots usually include one principal component on the x-axis and another principle component on the y-axis. If these PCs were successful in minimizing noise and identifying variance in spectra that indicate an underlying pattern or structure, there should be clustering of samples on the scores plot. In a simulated example dataset of seven objects measured at 12 wavelengths shown in Figure 4A, a scores plot was created to visualize these patterns. Figure 4B shows the scores plot for the data, and the data points indicate two clusters A and B along with a gradient in between. Although pure groupings of A and B may be separate, gradients may form which indicate properties that lie on a range between the two clusters.

The loadings can be analyzed through two different plotting techniques. One is identical to the scores plot but has the loadings of two different principal components plotted against each other. Another option is plotting the loading values individually as a function of wavenumber alongside a sample spectra. As shown by in Figure 4C, this visualization shows the similarities between the loading variance and the original spectra in Part A. It is important to note that loadings are in fact not spectra. Loadings are representative of the variance between all spectra and not any individual spectrum. This arises from the orthogonality of each PC; where PC 1 represents an “average loading,” PC 2 represents the “contrast loading” or the difference in absorption peaks and PC1.[17]

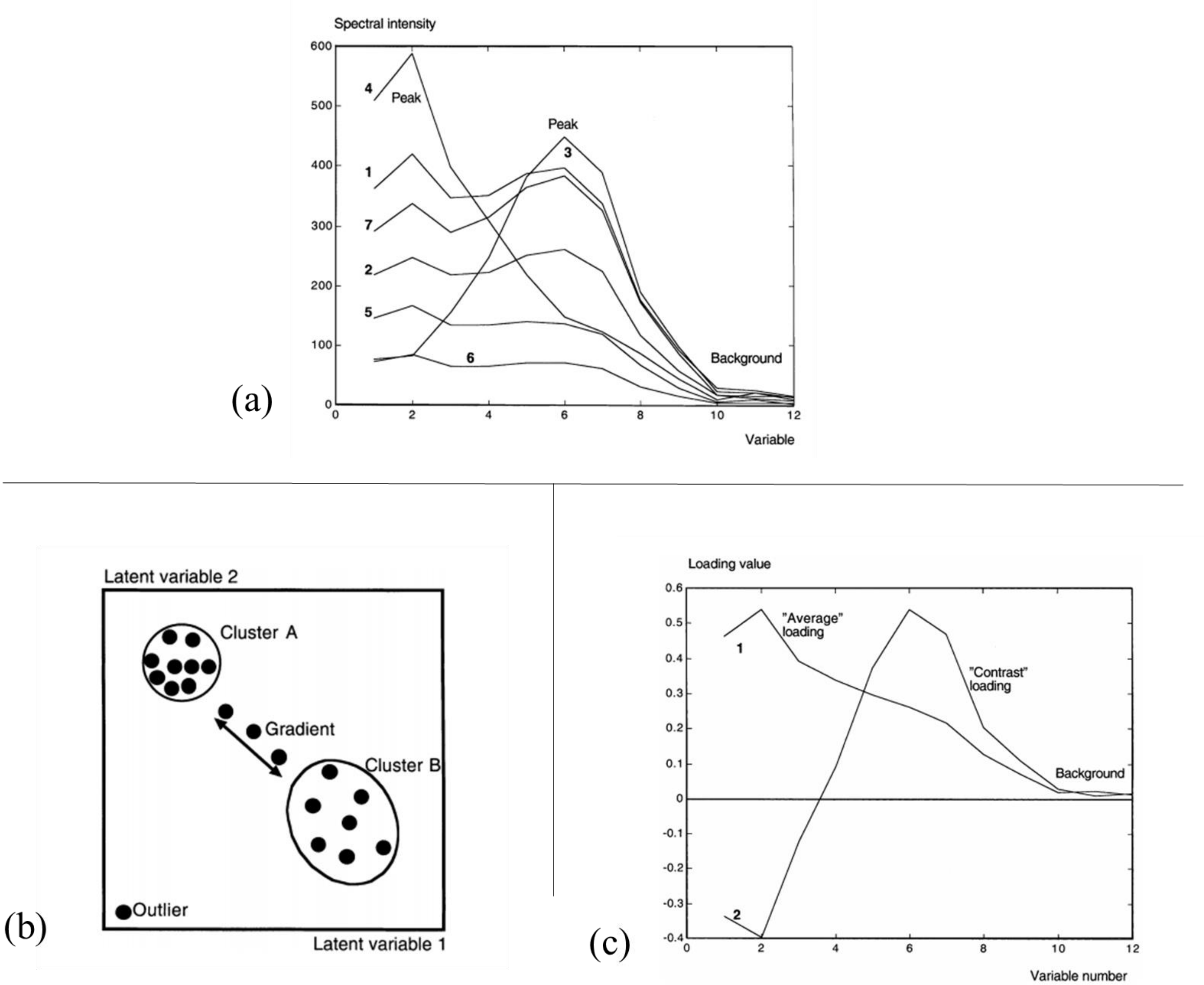


Figure 4: This data is a simulated set of 7 samples that include the spectral intensity at 12 different wavelengths, as shown by Part A. Part B shows the scores plot for PCs 1 and 2 for a theoretical dataset. Part C shows the loadings of PCs 1 and 2 for the dataset.[17]

2. Experimental

2.1 Fabrication of $PbTiO_3$ mesoporous film

A sol gel solution of lead acetate and titanium isopropoxide in ethanol with ethyl cellulose as the polymer was used. The FTO substrate was cleaned with by sonication in a series of distilled water, acetone, isopropanol, and ethanol. Films were made by placing and spreading 1 drop of the precursor sol gel solution across a fluorine doped tin oxide (FTO) glass substrate. All films were then annealed for 60 minutes at 600°C and were gradually brought to room temperature over 20 minutes. To minimize any variance during film preparation or measurements, all 44 $PbTiO_3$ films were created in the same day.

2.2 Dye loading of substrates

Groupings of four annealed films were randomly assigned a dye loading time (1, 5, 10, 15, 30, 45, 60, 90, 120 minutes and 24 hours) using a randomized number generator in R to account for any variations in films that might have occurred during the preparation and annealing process. Four additional films were assigned to be dye loaded for 1 minute the following day to control for any unexpected changes in the dye loading environment that could occur over 24 hours as a check for the 24 hour dye loading time group. Once the films completed their designated time submerged in P1 solution, they were rinsed with acetonitrile and dried in regular atmospheric conditions.

2.3 Spectroscopic Measurements

The $PbTiO_3$ films were measured by the Bruker ALPHA FTIR spectrometer using OPUS spectroscopy software version 7.2.139.1294 with an ATR attachment. To make up one spectra measurement, 24 scans were done at a spectral resolution of 4 cm^{-1} in absorbance mode in the range of 400 – 4000 cm^{-1}. The entire film on the FTO glass substrate was used for measurement, instead of scraping the film into a powder. UV-Vis spectra was collected using a Varian Cary 5000 UV-Vis-NIR spectrophotometer with Varian computer software.

2.4 PCA Calculations

PCA analysis was done using the ChemoSpec package in R created by Bryan Hanson.[18] This included the importing, cleaning, plotting, and creating the principal components. A sample flow of the coding used is provided in the Appendix.

2.5 DFT Calculations

Gaussian version 5.0.9 was used to build the P1 molecule and set up the DFT calculation. The calculation was done on a pure P1 molecule with no additions or changes to its original structure.

3. Results and Discussion

3.1 Dye loading time variance

For this experiment, we use PCA as a method to cluster FTIR spectra to probe any changes in the P1-$PbTiO_3$ binding mode with respect to dye loading time. Preliminary research indicates a difference in performance predicated upon dye loading times, as shown in Figure 5. Although the fill factor for each dye loading time is similar, there seems to be two groupings based on the photocurrent density and open circuit voltage. DSSCs that were dye loaded for 15 and 30 minutes had a higher in J_{sc} comparison to 60, 120, and 240 minutes and 24 hours. As previously mentioned, binding modes affect the efficiency of charge transfer and recombination which directly affects the J_{sc}, V_{oc}, and *ff*. Therefore the changes in dye loading times that give rise to variance in photoelectric conversion efficiency may indicate differences in binding modes.

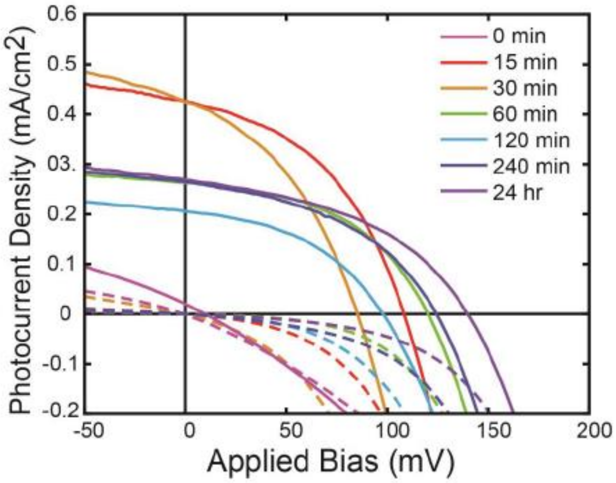


Figure 5: Graph of device performance for a set of $PbTiO_3$ films that were submerged in a P1 solution for varied times. The full lines depict the light *JV* curves whereas the dashed lines are the dark *JV* curves. Pink denotes a film without any P1 chromophore, red is dye loaded for 15 minutes, orange for 30 min, green for 60 min, light blue for 120 min, dark blue for 240 min and purple for 24 hours.

3.2 Theoretical values of P1 spectra

To better inform the presence of P1 in the experimental dataset, a DFT calculation was done for the chromophore. Figure 6 shows the spectra modeled for a pure P1 molecule from 0 to 400 cm^{-1}. A majority of the signal necessary to identify P1 is concentrated within the range of 1000 to 2000 cm^{-1}, so it is advantageous to use this range the entirety of the experiment. In addition, the FTIR in use tends to have inconsistent measurements at extremely low and extremely high wavenumber values. Reducing the range also reduces the amount of impact the noise from each end of the spectrum has on data analysis. By removing the fringes of a spectrum, the final data analysis will be less affected by data inconsistencies, outliers, and errors in the spectrometer measurements.

Visualizations of DFT models from the Gaussview software in conjunction with research on well-studied vibrational modes of organic molecules provided enough detail to assign modes to each peak in the DFT calculation. Table 1 summarizes these vibrational modes alongside their analogous peak. In particular, it is important to note the location of signals that correspond to carboxylic acid. Because it is the anchoring group of the chromophore, this functional group provides the most important peaks in the spectra and is also likely to shift or change in experimental data. The biding mode that adheres the molecule to the lead titanate surface can change the existing configuration of the anchoring group. The areas of focus along the spectrum for this group is the C-O stretch at 1120 cm^{-1}, the hydroxyl bend at 1190 cm^{-1}, and carbonyl stretch at 1820 cm^{-1}.

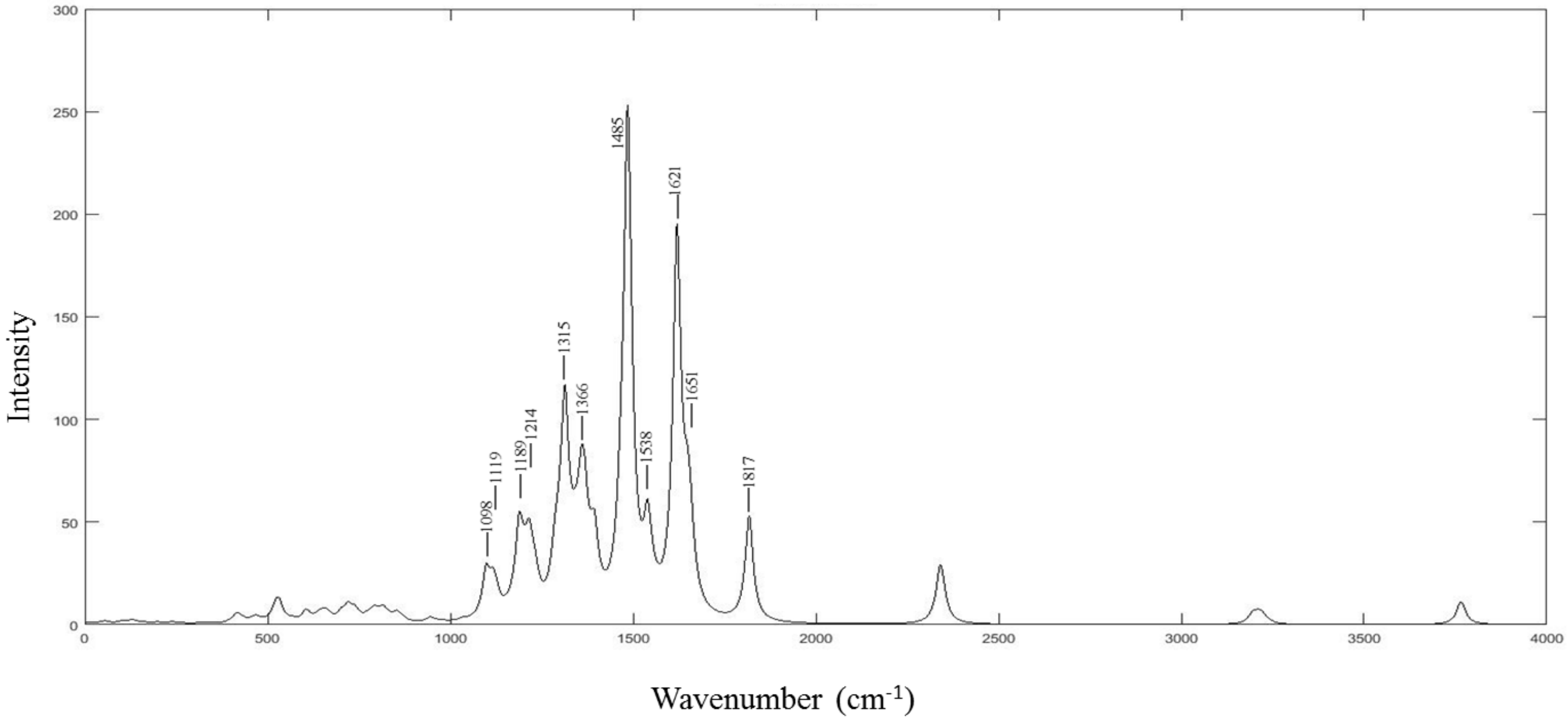


Figure 6: DFT calculation for a P1 chromophore using Gaussview. Peaks discussed in Table 1 are labeled with their exact wavenumber for reference.

Table 1: Peak labeling of DFT calculation in Figure 6. Each vibrational mode corresponds to a peak located by the specified wavenumber.

Vibrational Mode	Wavenumber (cm^{-1})
Carboxylic acid	
C-O stretch	1119
O-H bend	1189
C=O stretch	1817
Triphenyl amine	
C-H bend	1119, 1189, 1214
C-N stretch	1189, 1315, 1366
C=C stretch	1485, 1538, 1651
Thiophene	
C-H bend	1098
C=C stretch	1485, 1538
Malonitrile	
C=C stretch	1621

Although DFT has a lower computational cost than other modeling methods, there are many assumptions made that may not fit the experimental data. To better understand the limitations of the theoretical model, Figure 7 shows the DFT calculation graphed alongside a representative sample from each of the two datasets that will be described in later sections.

In comparing the three spectra, there are a number of similarities and differences. It is clear that each contain three prominent peaks in around the 1300-1650 cm^{-1} range which represent stretching primarily in the triphenyl amine but also in the malonitrile groups. In fact, the DFT peak at 1315 cm^{-1} is fairly consistent with the experimental data. However, the two other peaks that were predicted to be located close to 1485 cm^{-1} and 1621 cm^{-1} have comparatively lower wavenumber and are instead shifted left on the plot. Shifting in these peaks along with broadening can come from numerous sources, including the resolution of the spectrometer being used and experimental conditions not modeled in the DFT calculation. Other than these three peaks, it becomes increasingly challenging to prove a relationship between experimental values and the theoretical.

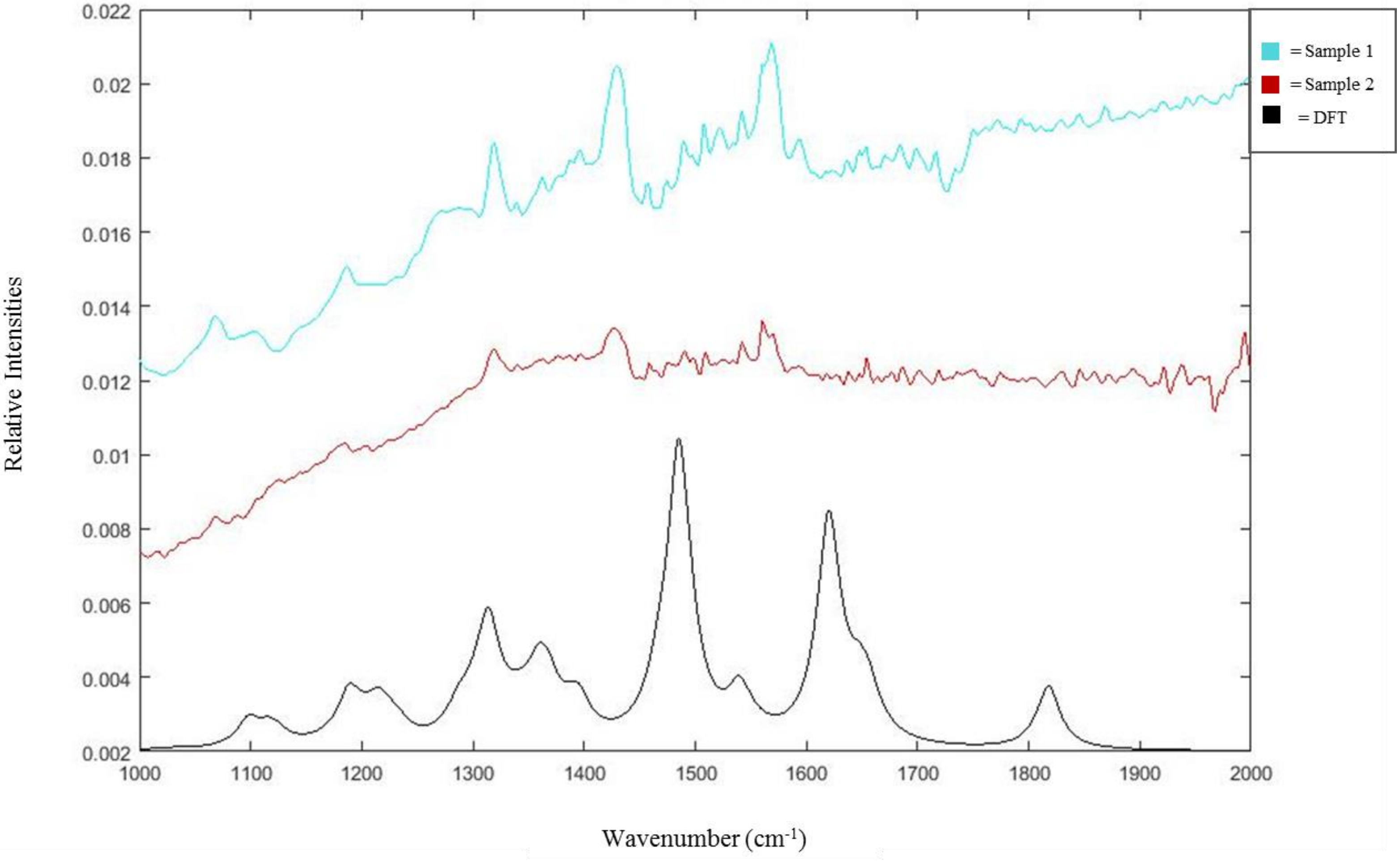


Figure 7: Relative intensities of experimental FTIR data versus the DFT calculation of P1. Sample 1 comes from the dataset discussed in section 3.4 and was dye loaded for 90 minutes, whereas sample 2 was dye loaded for 60 minutes and is pulled from the dataset discussed in section 3.3.

3.3 Computing presence of P1 from initial set of experimental spectra

To test the robustness of the PCA model in indicating differences in spectra, it must first be able to distinguish dye-loaded films from non-dye-loaded films. Figure 8A shows a dataset of

$PbTiO_3$ films between 1000 and 2000 cm^{-1} with red spectra indicating dye-loaded films for 1 hour and blue spectra contain no trace of P1. Differences between the two types of films becomes more apparent in the processed spectra shown in Figure 8B, which is where the triphenyl amine peaks become more prominent in films that contain P1. The non-dye-loaded films lack these indicators.

The scores plot with the clearest separation of dye-loaded films from non-dye-loaded films (Figure 8C) includes principal components 1 and 2. These two principal components combined make up 86% of the variation in the dataset. The variation that these principal components represent is shown in Figure 8D.

By rendering clear groupings in the scores plot, principal component analysis is capable of identifying the presence of the P1 chromophore in FTIR spectra. However, this plot displays groupings that overlap with one another—the data points form that of a gradient pattern rather than two separate clusters. This overlap may be from materials that the substrates share such as lead titanate.

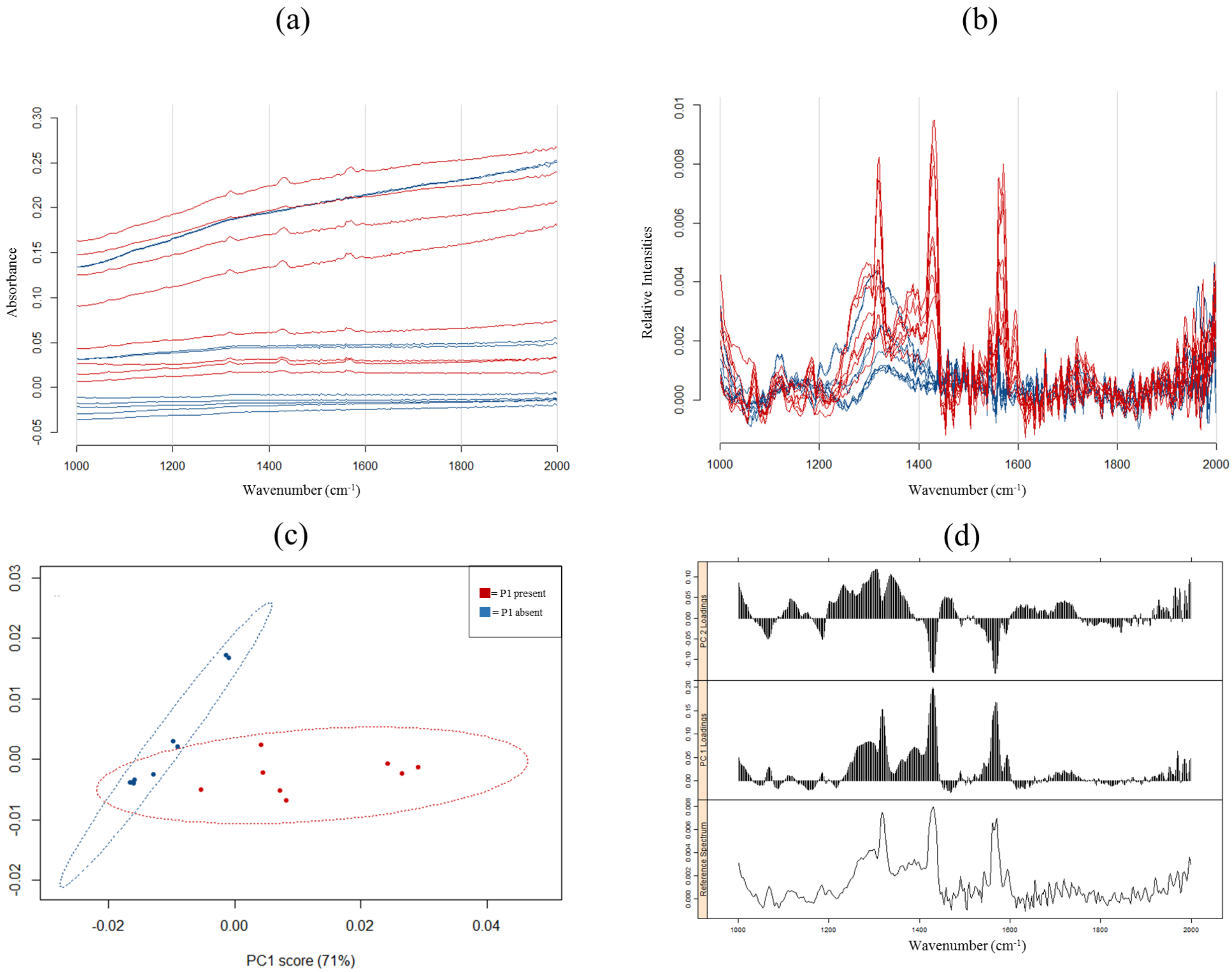


Figure 8: Analysis of dataset consisting of 17 FTIR measurements from 4 $PbTiO_3$ substrates, half of which were dye loaded in P1 for 1 hour. All spectra and data points colored red represent dye-loaded films whereas those in blue represent non-dye-loaded films. Part A shows the absorbance values for the samples from 1000-2000 cm^{-1} in their original, unprocessed form. Part B shows the same set of data with the baseline correction. Part C shows the scores plot for principal components 1 and 2. Part D shows the loadings plot for principal components 1 and 2 along with a reference spectrum, which is a selected spectrum from the dataset used as an example. In this plot, a dye-loaded film was used as the reference spectrum.

3.4 PCA of numerous dye loading times

Because the PCA model is able to identify the presence of P1, the next test of its robustness is in a dataset consisting of 44 spectra from films of varying dye loading times. Figure 9A shows the experimental dataset with the same wavenumber range that was used in the preliminary study. With the increased variance of dye loading times, the differences in spectra between the categories has come drastically unclear—there is little to no assumptions that can be made about the differences in dye loading times and bonding by simply looking at the graph.

To reduce noise, two inconsistent spectra and the ALS baseline correction technique were used. These preprocessing methods rendered the spectra shown in Figure 9B. This enhanced the signals found in the dataset, and the three peaks seem comparatively similar to the processed data

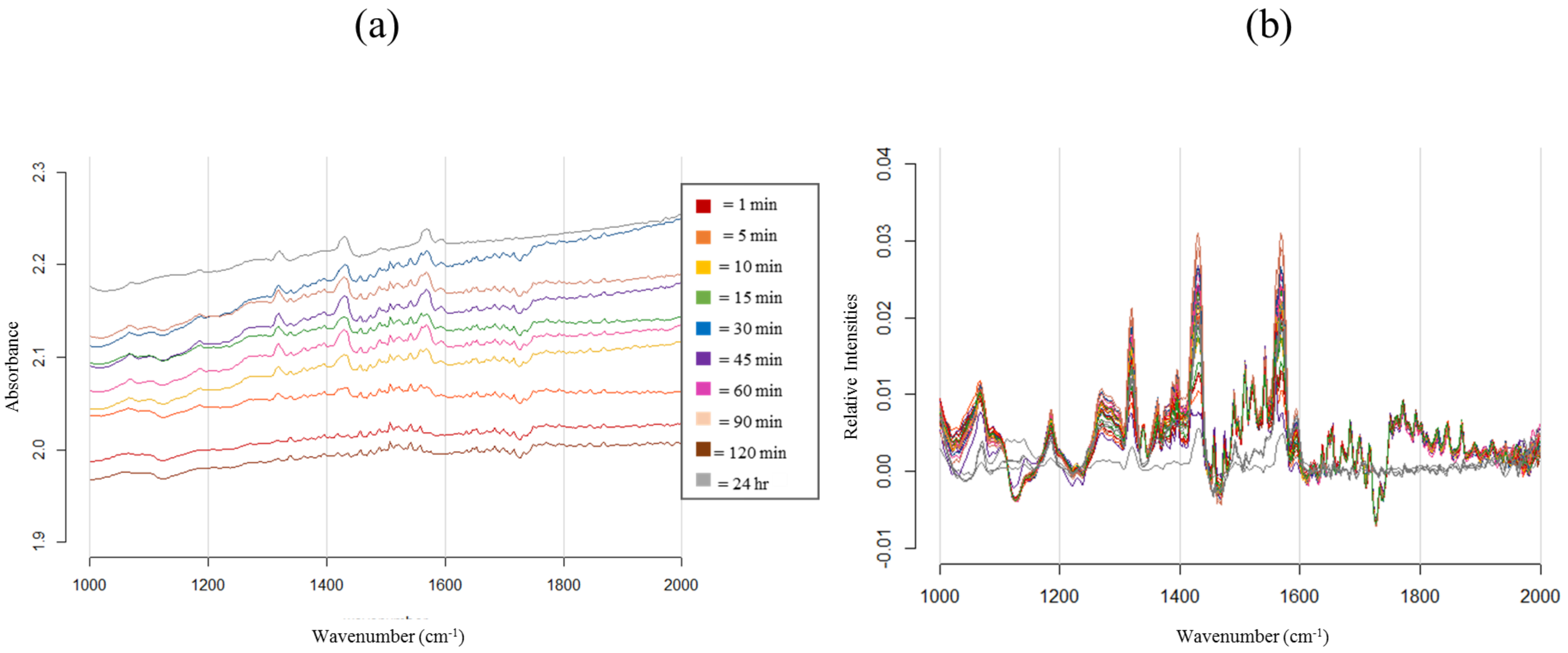


Figure 9: Part A shows the unprocessed FTIR spectra of 44 samples with varying dye loading times indicated in the legend included in Part B. The spectra shown in Part B do not include three outlier spectra which were in the 1 minute, 90 minute, and 120 minute dye loading time groups respectively. The ALS baseline correction technique was then used to render the final result.

in Figure 8B. However, unlike the clear difference between dye-loaded and non-dye-loaded spectra in Figure 8B, the outlier removal and baseline correction did not create clear visual differences in the spectra based on dye loading time.

It is clear from the two scores plots in Figure 10 that the PCA analysis was not successful in grouping time variance based on their binding modes. The scores plot of principal components 1 and 2 in Figure 10A indicates separation only of films that were dye loaded for 24 hours, which is most likely a result of a change in atmospheric conditions or measuring adjustments made by the spectrometer rather than an inherent change in the binding interactions of the chromophore-semiconductor interface. The drastic separation of the 24 hour dye loading time is only apparent in scores plots that compare principal component 1 against any subsequent component. This means the information that indicates a change in the atmospheric conditions is included within PC1. Based on the loadings plot in Figure 10B, the primary data variance that causes this is most likely the water vapor at 1750 cm^{-1} indicated by the red arrow in the loadings plot. Because of the significant changes the 24 hour dye loading time causes in the data, all spectra in this time period were removed in the scores plot for Figure 10C.

Figure 10C shows a different scores plot of principal components 2 and 5, and is representative of the majority of the remaining scores plot iterations based on two factors: there is no distinct separation of any time groupings and there is no indication of a gradient based on increasing or decreasing dye loading time. Although this scores plot does indicate areas where there are small signs of clustering, such as the ellipses encompassing films dye loaded 30 minutes, there are no definitive characteristics or patterns that can indicate information on the binding mechanism of P1. To determine whether a broader trend exists, this data was also grouped into low, medium, and high dye loading times located in Figure 11. The scores plots for part A and B in this figure still show no signs of clusters or a gradient.

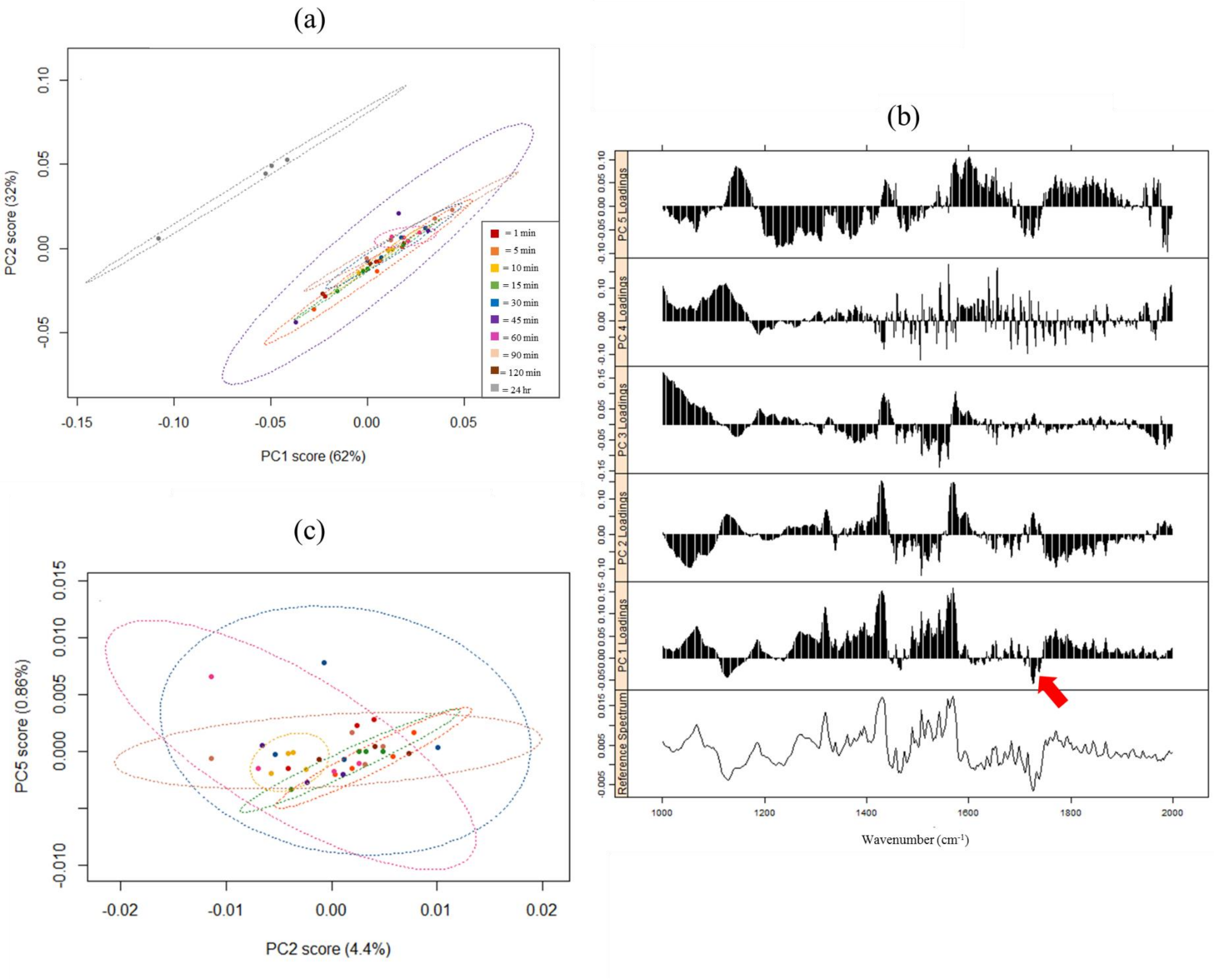


Figure 10: A) Scores plot for principal components 1 and 2 along with demonstrating a separation between the 24 hour dye loading time and the rest of the data. B) Loading plot for the PCA of this dataset and includes the loadings of principal components 1 through 7. C) Scores plot for principal components 2 and 5.

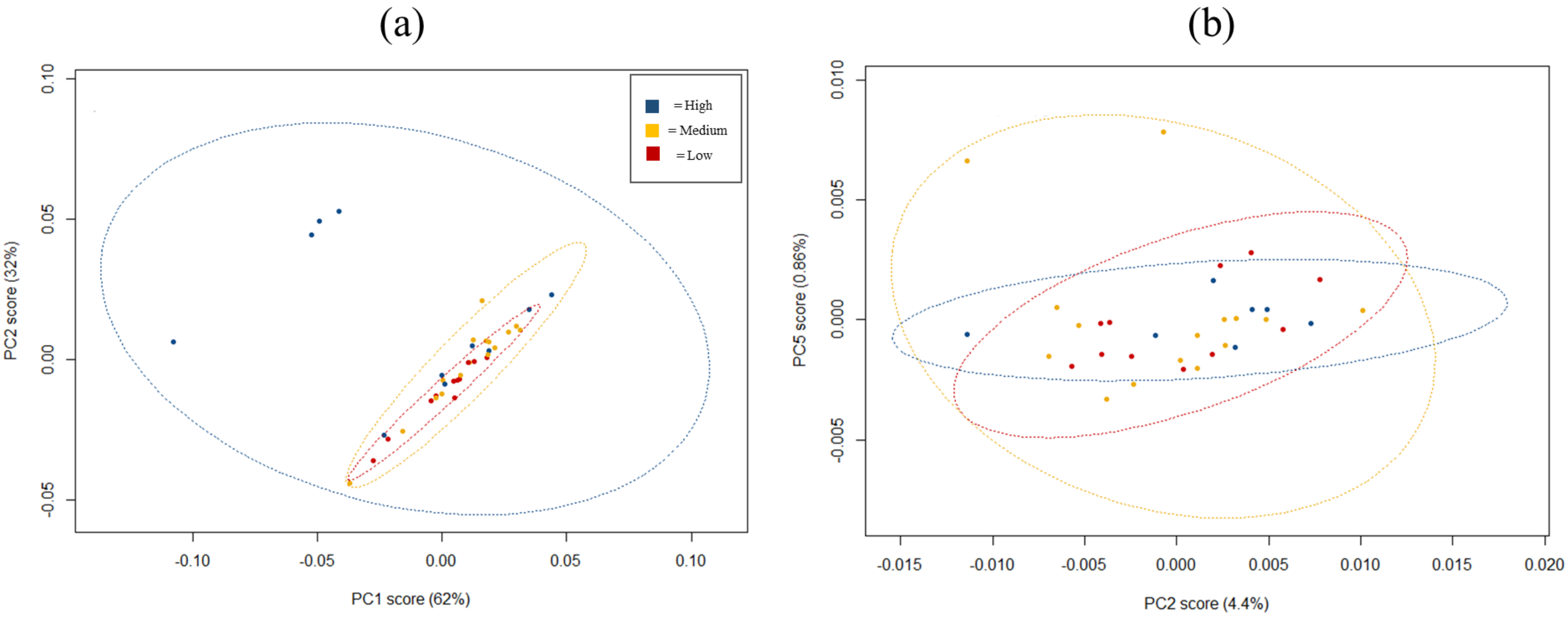


Figure 11: Part A and B are identical to Parts A and C in Figure 10 respectively, but the data points are clustered into high, medium, and low dye loading times. “High” represents loading times of 90 minutes, 120 minutes, and 24 hours. “Medium” represents dye loading times of 15, 30, 45, and 60 minutes. “Low” is dye loading times for 1, 5, and 10 minutes.

3.5 UV-Vis data and dye aggregation

Once the FTIR spectra was collected for the 44 films, they were then measured with a UV-Vis spectrophotometer. Figure 12A shows the spectra for each sample which are grouped by low, medium, and high dye loading times. Because there are no outlier spectra or noisy signals in the data, no data preprocessing was required. The scores plot of principal components 1 and 2 in Figure 12B show signs of a gradient from low to high dye loading times.

Because PCA was capable of separating UV-Vis data based on dye loading times, it is possible that dye aggregation and/or noisy FTIR spectra may be primary hindrances in this experiment. With regards to dye aggregation, this would be true in the case where the P1 solution that lead titanate films are immersed in is too concentrated which creates a surplus of chromophores onto the surface. Instead of measuring spectra for films that may have different binding modes predicated upon dye loading time, this experiment measured the aggregation and collection of dye molecules layering on top of one another. Thus the sample itself is so oversaturated with P1 molecules that it squelches the signal of the P1-$PbTiO_3$ binding mode. For noisy FTIR spectra, when comparing UV-Vis data (Figure 12A) to FTIR data (Figure 9A) there seems to be clear differences in the noisiness of the spectra collected. Although the data preprocessing led to a normal distribution in data variance for the FTIR spectra, this may not have been successful in eliminating noise and emphasizing information in the dataset. Therefore, PCA separation does not occur.

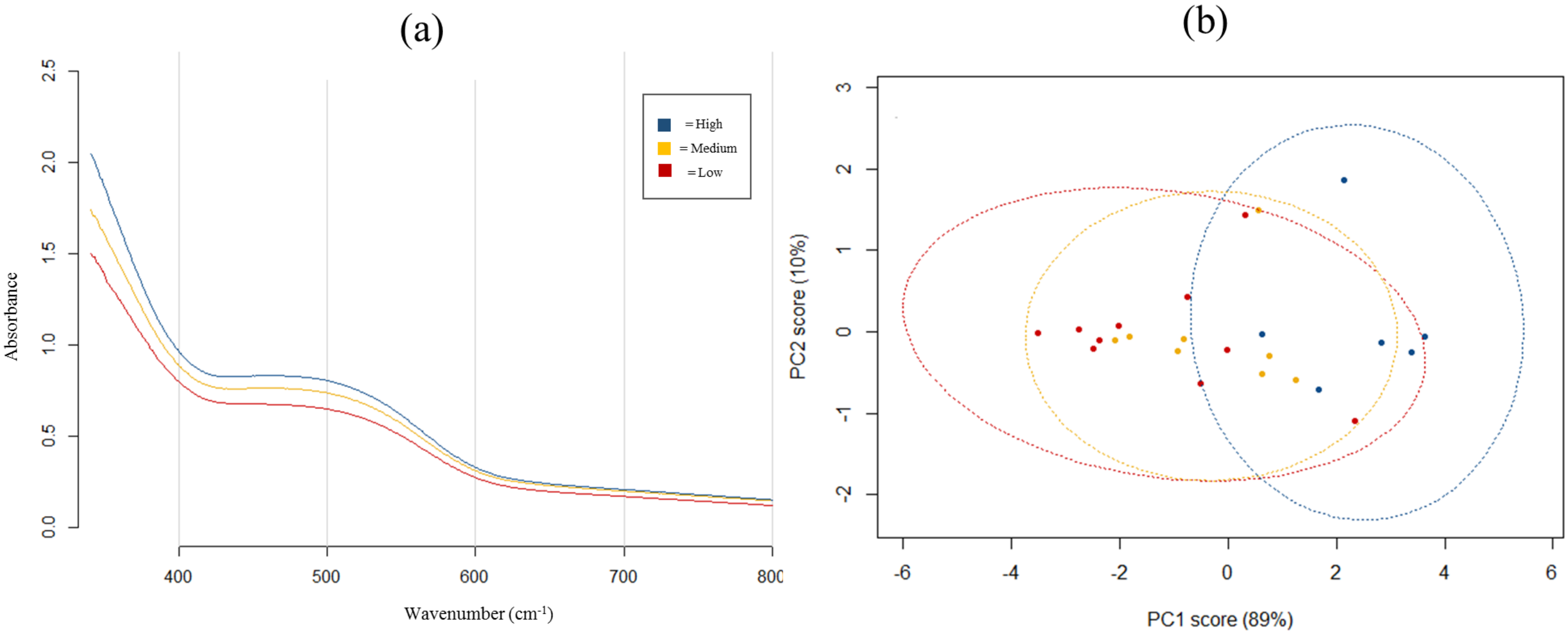


Figure 12: A) Three representative spectra from UV-Vis data, where “high” represents loading times of 90 minutes, 120 minutes, and 24 hours. “Medium” represents dye loading times of 15, 30, 45, and 60 minutes. “Low” is dye loading times for 1, 5, and 10 minutes.. B) Scores plot of UV-Vis spectra using same grouping.

4. Conclusion

This experiment presented principal component analysis as a method to identify binding modes at the P1-$PbTiO_3$ interface. In theory, the incorporation of multivariate statistics into conventional data analysis should render a better understanding of the information within these large datasets. In this case it does not. Although the scores plots of principal components from the FTIR data were capable of distinguishing the presence of P1 on the semiconductor surface, a variance of dye loading times did not result in clusters of data points. What did result, however, was a gradient from low to high dye loading times of UV-Vis data. This may indicate an aggregation of P1 that overpowers the P1-PbTiO3 binding modes or FTIR data with too low of a signal-to-noise, or a combination of both. Further research should be done on the optimal concentration of P1 solution, its optimal dye loading time for lead titanate films, and reducing noise in FTIR spectra both through experimentation and data preprocessing.

Despite the setbacks seen in the data analysis of this experiment, this should not hinder further study into PCA as a tool for understanding binding modes for DSSCs. In fact, exploration into this area should be increased. PCA is a broad technique used in many fields and can therefore be translated to a number of alternate chromophores and semiconductors.

Appendix

Sample flow of ChemoSpec analysis in R:

```
library(R.utils)

library(ChemoSpec)

setwd("C:/Users/zemaitis/Documents/UNC/Cahoon Lab/R/FTIR - PbTiO3 09212016")

files2SpectraObject(gr.crit = c("min01", "min05", "min10", "min15", "min30", "min45",
"min60", "min90", "min120", "hr24"), gr.cols = ("red3", "orangered", "darkgoldenrod2",
"forestgreen", "dodgerblue4", "purple4", "violetred2", "salmon3", "orangered4", "gray48"),
freq.unit = "wavenumber",int.unit = "%transmittance", out.file = "times", sep = ",")

timevariance<-loadObject("times.RData")

timevariance <- binSpectra(timevariance, bin.ratio = 2)

plotSpectra(timevariance,main="Time Variance Raw",which = 1:51,xlim = c(500,2500),yrange
= c(0.58,1.1), offset = 0, lab.pos = 7000)

class<-c_pcaSpectra(timevariance, choice = "noscale")

ChemoSpec::plotScores(timevariance, main = "Time Variances for Current Data", class,pcs =
c(2,4), ellipse = "none", tol = 0.01, leg = "none")
```

References

1. O'Regan, B. & Grätzel, M. A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO2 films. *Nature* **1991**, 353, 737–740.
2. Odobel, F., Pellegrin, Y., Gibson, E.A., Hagfeldt, A., Smeigh, A.L. and Hammarström, L., 2012. Recent advances and future directions to optimize the performances of p-type dye-sensitized solar cells. *Coordination Chemistry Reviews* **2012**, 256, 2414–2423.
3. National Renewable Energy Laboratory. Best Research-Cell Efficiencies. **2017**.
4. Hagfeldt, A., Grätzel, M. Light induced redox reactions in nanocrystalline systems. *Chem. Rev.* **1995**, 95, 49–68.
5. Lee, K.E., Gomez, M.A., Elouatik, S. & Demopoulos, G.P. Further understanding of the adsorption mechanism of N719 sensitizer on anatase TiO2 films for DSSC applications using vibrational spectroscopy and confocal Raman imaging. *Langmuir* **2010**, 26(12), 9575–9583.
6. Li, L., Gibson, E.A., Qin, P., Boschloo, G., Gorlov, M., Hagfeldt, A. and Sun, L. Double-Layered NiO Photocathodes for p-Type DSSCs with Record IPCE. *Adv. Mater.* **2010**, 22, 1759–1762.
7. Shi, Z., Lu, H., Liu, Q., Deng, K., Xu, L., Zou, R., Hu, J., Bando, Y., Golberg, D. and Li, L. NiCo2O4 Nanostructures as a Promising Alternative for NiO Photocathodes in p-Type Dye-Sensitized Solar Cells with High Efficiency. *Energy Technol.* **2014***,* 2, 517 – 521
8. Qin, P., Zhu, H., Edvinsson, T., Boschloo, G., Hagfeldt, A. & Sun, L. Design of an Organic Chromophore for P-Type Dye-Sensitized Solar Cells. J. Am. Chem. Soc. **2008**, 130-27, 8570–8571.

9. O'Rourke, C. and Bowler, D.R. DSSC anchoring groups: a surface dependent decision. *Journal of Physics: Condensed Matter* **2014**, *26*(19), 195302.
10. Engel, T., Reid, P. *Physical Chemistry.* Second Edition. Upper Saddle River (NJ): Pearson Education, Inc. **2006**.
11. Lasch, P. and Naumann, D. Infrared spectroscopy in microbiology. *Encyclopedia of analytical chemistry* **2015**.
12. Rinnan, Å., van den Berg, F. and Engelsen, S.B. Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry* **2009**, *28*(10), 1201-1222.
13. Eilers, P.H. and Boelens, H.F. Baseline correction with asymmetric least squares smoothing. *Leiden University Medical Centre Report*, **2005.** *1*.
14. Adam, C.D., Sherratt, S.L. and Zholobenko, V.L. Classification and individualisation of black ballpoint pen inks using principal component analysis of UV–vis absorption spectra. *Forensic science international* **2008**. 174(1), 16-25.
15. Cangelosi, R. and Goriely, A., Component retention in principal component analysis with application to cDNA microarray data. *Biology direct* **2007**, *2*(1), 2.
16. Wold, S., Esbensen, K. and Geladi, P., 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* **1987**, *2*(1-3), 37-52.
17. Geladi, P. Chemometrics in spectroscopy. Part 1. Classical chemometrics. *Spectrochimica Acta Part B: Atomic Spectroscopy* **2003**, *58*(5), 767-782.
18. Hanson, B.A. Chemospec: an R package for chemometric analysis of spectroscopic data and chromatograms. *Package Version* **2015**, 2(2).