Application of Infrared Spectroscopy and Partial Least Squares Discriminant Analysis to Determine the Gonotrophic Stage of *Aedes triseriatus*

A thesis presented to the faculty of the Graduate School of Western Carolina University in partial fulfillment of the requirements for the degree of Masters of Science in Chemistry.

By

Mark Anthony Rothermund

Advisor: Dr. Scott W. Huffman
Associate Professor of Chemistry
Department of Chemistry & Physics

Committee Members: Dr. Carmen L. Huffman, Chemistry & Physics
Dr. Brian D. Byrd, Health and Human Sciences

April 2022

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

FT	Fourier Transform
FT-IR	Fourier Transform Infrared
HDF	Hierarchical Data Format
IR	Infrared
JDX	The Joint Committee on Atomic and Molecular Physical Data - Data Exchange
MIDS	Mid-Infrared Spectroscopy
PCA	Principal Component Analysis
PLS	Partial Least Squares
PLS-DA	Partial Least Squares Discriminant Analysis
SG	Savitzky–Golay
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate

ABSTRACT

APPLICATION OF INFRARED SPECTROSCOPY AND PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS TO DETERMINE THE GONOTROPHIC STAGE OF *AEDES TRISERIATUS*

Mark Anthony Rothermund, M.S., Chemistry

Western Carolina University (April 2022)

Advisor: Dr. Scott W. Huffman

Mosquitoes are among the deadliest creatures in the world due to their propensity for spreading pathogens to humans. Surveillance is an important step in controlling and monitoring mosquito populations. The most common technique used for mosquito surveillance requires a highly trained entomologist to identify and determine information such as sex, gonotrophic stage, infection status of the mosquitoes morphologically by means of dissection. Identification using this method is time-consuming and requires skills that only highly trained entomologists possess, which limit the sample size of tested mosquitoes. A new method using Fourier transform infrared (FT-IR) microspectroscopy eliminates these restrictions by streamlining mosquito sample processing and lowering the skill required to perform the method. The method outlined in this study can be completed more quickly and by technicians of varied skill levels. Samples of parous and nulliparous *Aedes triseriatus* (170 samples) identified by a trained entomologist were used to test the method's ability to discriminate parity. Mid-infrared spectra of the mosquitoes were collected, preprocessed, and partial least squares discriminant analysis (PLS-DA) was used to discriminate between the parous and nulliparous mosquitoes. The method identified the parity status of the mosquitoes with 100% accuracy, 100% true positive rate (TPR), and 100% true negative rate (TNR).

CHAPTER ONE: INTRODUCTION

Background

Motivation

Mosquitoes are among the deadliest creatures in the world, causing considerable morbidity, mortality, and economic strain.[1] More than 80% of the global population is at risk of vector-borne disease, and mosquito-borne diseases contribute the most to human vector-borne disease burden.[2] Mosquito-borne diseases include dengue fever, Zika virus disease, West Nile fever, malaria, chikungunya, St. Louis encephalitis, yellow fever, and encephalitides such as Jamestown Canyon encephalitis and La Crosse encephalitis that are caused by California serogroup viruses.[1] The emergence and resurgence of vector-borne diseases are attributed to changes in public health policy, insecticide and drug resistance, shifts in emphasis from prevention to emergency response, demographic and societal changes, and genetic changes in pathogens.[2,3] Mosquito control efforts such as source reduction, larvicides, and adulticides remain a primary method of disease control and prevention.[4] Surveillance, an important step of mosquito control efforts, is needed to reduce nuisance and the spread of mosquito-borne diseases.[5] Currently, surveillance is commonly conducted by a skilled biologist using morphological identification.[5] Morphological surveillance is time-consuming and labor-intensive, and the current limitations involved with morphological surveillance leads to a limited sample size of identified and analyzed mosquitoes that may not represent the entire population of mosquitoes very well. A streamlined surveillance approach using infrared spectroscopy to analyze the mosquitoes' biochemistry would require only skills that should be ubiquitous for average laboratory technicians. Additionally, information such as sex, gonotrophic stage, blood meal status, and infection status are difficult, perhaps even impossible in some cases, to determine morphologically or anatomically without knowing the history of the mosquito; however, this information theoretically can be determined easily by analyzing the biochemical composition of the mosquito using infrared spectroscopy.

Importance of Parity Assessment

Most female mosquito species acquire blood meals to develop their eggs after mating.[6] In the acquisition of blood, pathogens may be transferred from the host to the mosquito wherein the mosquito becomes a vector that has the potential to spread the pathogen to another animal. Only parous mosquitoes are potentially infectious; therefore, distinguishing between mosquitoes that are parous (having produced offspring) and mosquitoes that are nulliparous (not having produced offspring) is an important aspect in mosquito control efforts. A shift in parity structure in a population toward a lower proportion of parous mosquitoes translates to a reduction in disease transmission.[6] The current standard technique of assessing mosquito parity requires a specialized biologist and involves a delicate dissection of the mosquito to inspect ovaries in order to evaluate their gonotrophic history.[7] Using infrared spectroscopy to assess parity, on the other hand, would require comparitively minimal skill.

Morphological Approach to Assessing Parity

Currently, parity is determined by dissection. After the maturation of the first batch of eggs, irreversible changes occur in the ovaries of the mosquitoes.[8] The tightly coiled tracheoles in nulliparous mosquito ovaries stretch and uncoil when a mosquito becomes parous as shown in Figure 1.[8] Entomologists are not only required to carefully dissect a mosquito to view its ovaries, but must also be able to differentiate between coiled and uncoiled tracheoles in mosquito ovaries to determine parity.

Infrared Spectroscopy

Infrared (IR) spectroscopy is an analytical technique that is rapid, non-invasive, reagent-free, highly sensitive, and simple to use. IR spectroscopy also provides information about the chemical composition of the sample, allowing for the ability to differentiate between proteins, conformations of the same protein, and even disease states in humans.[9,10] IR spectroscopy and other spectroscopic techniques share commonality in using the detection of light as a measurable signal. Like other spectroscopic techniques, the Beer-Lambert law can be used to determine the ab-

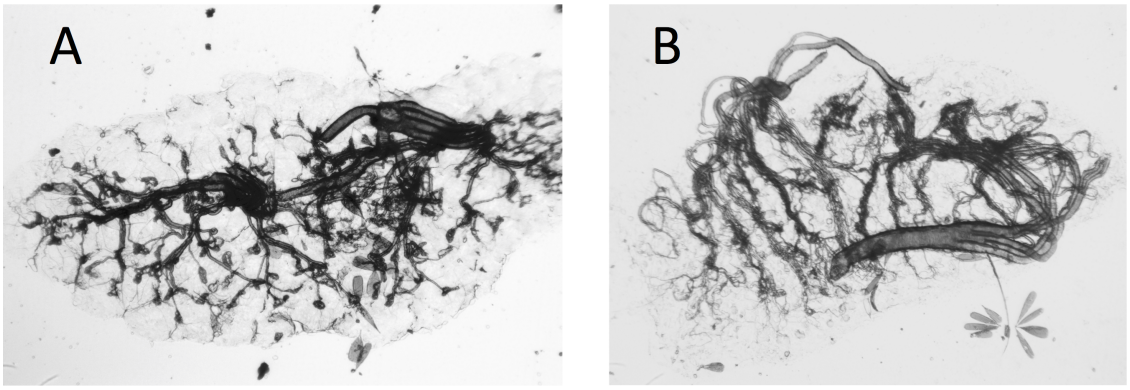

Figure 1. Comparison of coiled tracheoles in nulliparous *Culex quinquefasciatus* ovaries (A) and uncoiled tracheoles in parous *Culex quinquefasciatus* mosquito ovaries (B).[8]

sorbance (A) of a sample.[11] The absorbance of a sample is directly proportional to the pathlength (l) and the concentration (c) of the sample, as shown:

$$A = \epsilon c l \tag{1}$$

where ϵ represents molar absorptivity and is a constant dependent on the material of the sample. IR spectroscopy provides chemical information by irradiating a sample with IR light. In a Fourier transform infrared (FT-IR) spectrometer, all wavelengths of light in the specified interval are emitted simultaneously.[12] A Fourier transform (FT) algorithm allows the whole spectrum to be scanned at the same time through interferometric modulation by an interferometer.[12,13] The modulated beam exits the interferometer, which is most commonly a Michelson interferometer, and passes through the sample.[13] The light is then detected by the detector, and the interference pattern is converted to a digital signal that is then transformed into an FT-IR spectrum using Fourier transformation.[13] Mid-infrared spectroscopy (MIRS) uses wavenumbers within the region of 4000–400 cm^{-1}, and because most modern MIRS instruments use FT, FT-IR spectroscopy and MIRS are typically synonymous.[13]

FT-IR Microspectroscopy

To aid in reproducibility of the proposed method and to increase the throughput of mosquito processing, a special kind of IR spectroscopy method, Fourier transform infrared (FT-IR) microspectroscopy, can be utilized. FT-IR microspectroscopy combines the FT-IR spectrometer with a microscope that can focus on a specific area of the mosquito.[4] The addition of the microscope aids in reproducibility and consistency by allowing the technician to select the same anatomical position of the mosquito every time for each sample. More importantly, FT-IR microspectroscopy allows the production of spectra that represent the chemical makeup of the region being probed. When analyzing the spectra to discriminate betweeen parous and nulliparous mosquitoes, it is unlikely that there will be any obvious visible differences in the spectra. There will be several subtle differences in the spectra that can only be pinpointed by dimensionality reduction in chemometric techniques. Before chemometric techniques can be applied, the data must first be subjected to several preprocessing steps to prepare it for analysis.

Selection of Preprocessing Techniques

The importance of preprocessing prior to several chemometric methods, including PLS-DA, has been demonstrated in many different studies.[14–18] PLS-DA has been shown to perform better with preprocessing than without.[14] Each method of preprocessing has its own role in preparing data for chemometric analyses.

Outlier Detection and Removal

Failure to remove outliers negatively affects statistical analyses by skewing distribution of spectra.[15] Several different methods of detecting outliers exist, and they have their own utility in different disciplines and applications. Outliers in FT-IR can be detected and removed at the discretion of the researcher within reason. Reasons for spectral removal may include improper technique utilized during spectra collection and noticeable visual differences between the outlying spectra and the average spectrum. More outliers that may not be obvious to the researcher may be detected by means of chemometric techniques such as PCA.[19,20] Outliers determined by PCA can

be visualized as data that is found outside of the cluster.

Cropping

Cropping, although not utilized specifically for assisting processing by chemometric methods, helps the researcher understand what regions of the spectra are important for discrimination. Cropping is necessary so that further processing steps focus only on the region of the spectra with varying chemical components detected by IR that are useful for chemometric analyses. Chemometric techniques can be utilized on several spectral regions after cropping, and performance metrics of the chemometric technique determine the spectral region's utility in discrimination of the particular variable of study.

Normalization

The importance of normalizing data prior to PLS-DA has been demonstrated by Lee *et al.*[14] In their study exploring practical impacts of data processing methods in IR spectra using PLS-DA, normalization was listed as the second most important step of preprocessing.[14] Normalizing is necessary because chemometric algorithms such as PCA and PLS-DA do not make assumptions to correct maldistribution of data. Normalization serves to equalize the statistical weight of each sample and minimizes the effect of light scattering.[16]

Smoothing and Differentiation

Smoothing using the Savitzky–Golay (SG) algorithm serves two purposes: minimizing the effect of noise and correcting sloped baselines brought on by a scattering medium.[17] The SG algorithm has three parameters: window size, polynomial order, and derivative order. The window size is the number of data points used to calculate a function of best fit with a given polynomial order.[21] The function is then differentiated according to the given derivative order.[21] Differentiation is used to de-emphasize and minimize the effect of a non-flat baseline in fitting the spectra.

Data Analysis

To analyze the spectral data, different techniques of classification can be used depending on the required robustness. More robust techniques like PLS-DA have the benefit of yielding higher ac-

curacy in predictions; however, they require more data processing than simple wavelength comparisons.

Wavelength Comparisons

The least complex data analysis technique explored in this study is wavelength comparison. Wavelength comparison is the visual analysis of the absorbance in the spectra of parous mosquitoes at different wavelengths (which may manifest as peaks), and comparing with the absorbance in the spectra of nulliparous mosquitoes at the same corresponding wavelengths. Wavelength comparisons can be easily achieved in instances where a peak or band in the spectra exists only in nulliparous mosquitoes and does not exist in parous mosquitoes or vice-versa. Using wavelength comparisons, however, will likely not yield a conclusive result since using wavelength comparisons as a sole method of differentiation is used most effectively with spectra of pure compounds.[4] Because the spectra involved in this study are of biological samples that consist of complex mixtures, a more robust technique is required.

Chemometric Techniques

Principal component analysis (PCA) is a chemometric technique used in data reduction and exploratory analysis of high-dimensional data sets, such as spectra.[22,23] PCA works by identifying which dimensions in the datasets have the most variability and grouping the data based off that variability.[22] The most variable dimension becomes a new variable called the first principal component. Subsequent principal components are determined by spectral features of decreasing variability that are each orthogonal to each other.[22] Orthogonality of the new principal components is important in FT-IR spectra discrimination because FT-IR spectra original variables are highly covariant, and orthogonal principal components eliminates covariance and redundancy.[24] Another chemometric technique is Partial least squares discriminant analysis (PLS-DA). Like PCA, PLS-DA separates data into groups by using linear combinations of the original variables to create new latent variables, but the difference is that PCA is an unsupervised technique, and PLS-DA is a supervised technique that also determines the principal components differently.[23] PCA cre-

ates its principal components to maximize the variance in the dataset, whereas PLS-DA creates its latent variables to maximize the covariance between the training set and the class labels.

Unsupervised Versus Supervised Technique

An unsupervised technique is one in which the algorithm groups or clusters the data on its own, whereas in a supervised technique, the data is categorized by the algorithm into groups that are predetermined by the researcher.[23] There are advantages and disadvantages to each technique. PCA provides unbiased dimensionality reduction while PLS-DA is biased due to the added human component of the predetermined groups.[23] PCA and PLS-DA can be used in conjunction where the initial use of PCA can provide confirmation and an informative first look at the dataset structure prior to analysis by PLS-DA.[23]

PCA can be used to cluster the spectral data. If there are two clusters, one would correspond to parous mosquitoes, and the other would correspond to nulliparous mosquitoes. An accuracy score can be given to this method based on having the correct number of clusters, two, and correct discernment of the parity of the mosquitoes by the algorithm. If PCA performs poorly, PCA can be used as an outlier rejection tool, and the more robust PLS-DA can be used to discriminate the data. For PLS-DA, the data will be split into two sets, a training set and a validation set. The parous and nulliparous status of the mosquitoes as well as their spectra in the training set will be used to generate a PLS-DA model that can be used to predict the status of previously unfit spectra, the validation set.

PLS-DA Performance Metrics

Accuracy, true positive rate (TPR), true negative rate (TNR) are the metrics that are used to measure the PLS-DA model's ability to discriminate data in this study. In the binary classification system used in this study, the samples of mosquito were either actually POSITIVE for parity (parous) or NEGATIVE for parity (nulliparous). When the PLS-DA model classifies a validation input based on training data in a binary system, it can either predict each input as POSITIVE or NEGATIVE.[25] When the algorithm correctly predicts an actual POSITIVE input as POSITIVE, it

is counted as a true positive (TP). When the algorithm correctly identifies an actual NEGATIVE input as NEGATIVE, it is counted as a true negative (TN). The word TRUE in "true negative" refers to a correct prediction. Additionally, there are identifications of "false" positive (FP) and "false" negative (FN), which mean that the model predicted an actual NEGATIVE input to be POSITIVE, and an actual POSITIVE input to be NEGATIVE, respectively. The possible designations of TP, TN, FP, and FN from the binary system of POSITIVE or NEGATIVE determined by the researcher, known as the actual class, and the binary system of POSITIVE or NEGATIVE determined by the algorithm, known as the predicted class, can be represented in a 2×2 matrix known as a confusion matrix as demonstrated in Figure 2.[25]

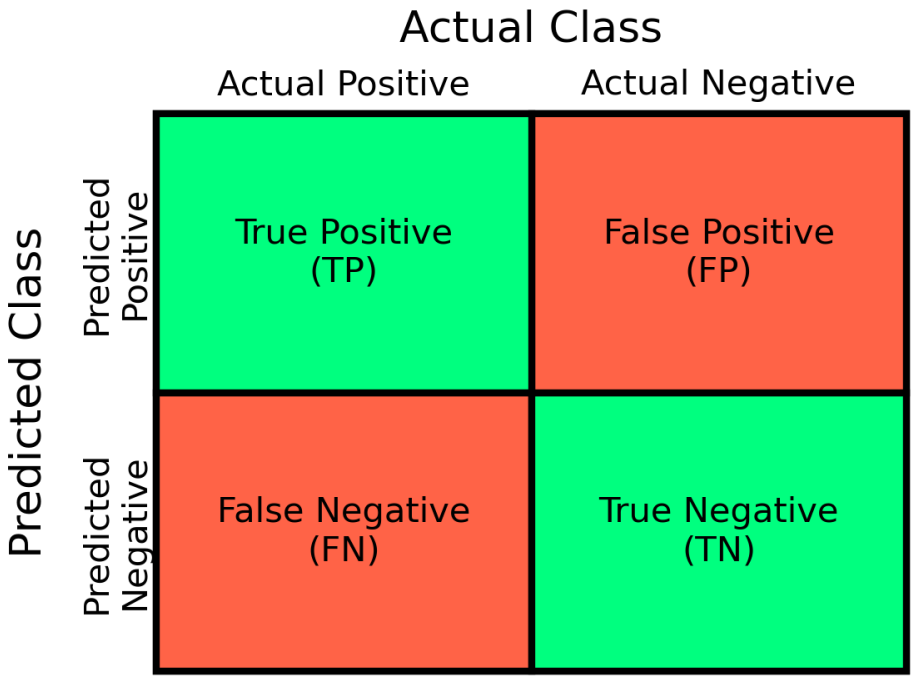

Figure 2. A 2×2 confusion matrix.

The accuracy of the model can be calculated with the following equation:[25]

$$Accuracy = \frac{n_{\text{TP}} + n_{\text{TN}}}{n_{\text{TP}} + n_{\text{TN}} + n_{\text{FP}} + n_{\text{FN}}} \cdot 100\% \tag{2}$$

where n_{TP} is the number of TP designations, n_{TN} is the number of TN designations, n_{FP} is the number of FP designations, and n_{FN} is the number of FN designations. The true positive rate (TPR) is an expression of a model's sensitivity.[25] TPR is calculated with the following equation.[25]

$$TPR = \frac{n_{\text{TP}}}{n_{\text{TP}} + n_{\text{FN}}} \cdot 100\% \tag{3}$$

The TPR takes into account only actual POSITIVE samples determined by the researcher, so the TPR can be thought of as the algorithm's accuracy in correctly identifying POSITIVE samples. The true negative rate (TNR) is an expression of the model's specificity and is calculated with the following equation.[25]

$$TNR = \frac{n_{\text{TN}}}{n_{\text{TN}} + n_{\text{FP}}} \cdot 100\% \tag{4}$$

The TNR takes into account only actual samples determined to be NEGATIVE by the researcher, so the TNR can be thought of as the algorithm's accuracy in correctly identifying NEGATIVE samples.

A model with accuracy, TPR, and TNR close to 100% indicates a good performance.

Hypothesis

The efficacy of FT-IR microspectroscopy combined with chemometric techniques has been demonstrated by Sroute *et al.*[4] in their study distinguishing between and classifying species of mosquitoes using spectral data coupled with PLS-DA. The hypothesis is FT-IR spectra coupled with PLS-DA can predict gonotrophic stage of *Aedes triseriatus* mosquitoes, the main vectors of La Crosse

virus which result in La Crosse encephalitis.[26]

CHAPTER TWO: EXPERIMENTAL

Sample Preparation

Colonized *Aedes triseriatus* (originally obtained from Michigan State University) were hatched, reared, and held using standard practices and conditions (27°C, 75% RH, 16:8 light:day photoperiod). Emergent adults (> 500) were placed in a single cage (length = 32 cm $\times$ width = 31 cm $\times$ height = 9 cm) and held for 7 days to allow for mating and maturation. Mosquitoes ($\approx$ 250) were removed from the original cage and placed in a secondary, identically sized cage; female mosquitoes were allowed to feed to repletion by providing a volunteer arm (B.D. Byrd) as a blood source; an oviposition substrate was provided 72 hours post feeding. Female mosquitoes were concurrently (same day) removed from both the original (nulliparous) and secondary (blood-fed) cages after cessation of oviposition. The mosquitoes were killed by freezing and stored at -20°C. A total of 170 mosquitoes were used for IR analysis; 99 were blood-fed (expected to be parous), and 71 had not blood-fed (known to be nulliparous).

Instrumentation and Setup Procedure

Samples were measured using the FT-IR microscope with the instrument parameters shown below in Table 1.

Before beginning spectrum collection, the instrument's dewar was filled with liquid nitrogen to cool the detector. With the microscope focused on a gold plate and the aperture completely open, collection was allowed to begin after the bench's interferogram stabilized.

Background Spectrum Collection Procedure

To collect a background spectrum, the microscope was focused on a gold plate with the aperture open. The background spectrum was collected as a single-beam spectrum. The background spectrum was saved as The Joint Committee on Atomic and Molecular Physical Data - Data Exchange (JDX) file format and opened in the OMNIC™ software as a background spectrum before collecting a sample spectrum.

Table 1. Key instrument parameters.

Parameter	Value
Microscope	Nicolet™ Centaurμs™
FT-IR Spectrometer	Nicolet™ IS™ 10
Software	OMNIC™ version 9.8.372
Wavenumber interval	4000–650 cm^{-1}
Near/mid/far IR	Mid
Detector	MCT/A, Liquid Nitrogen Cooled
Beamsplitter	KBr
Blank	Air
Scans	64
Resolution	4 cm^{-1}

Sample Spectrum Collection Procedure

To collect a sample spectrum, the center portion of the tibia of the hind leg of the mosquito shown in Figure 3 was measured. The tibia was chosen as the anatomical part for examination because it is easy for a technician with limited entomology training to identify. The mosquito's leg was placed under the microscope, and the microscope's coarse focus was adjusted to focus on the leg. Depending on the ability or inability of the microscope to focus on the leg, the tibia may have been separated from the rest of the leg to encourage it to lay flat on the microscope's stage. The microscope's aperture was then closed in on the center portion of the tibia. The fine focus of the microscope was adjusted to refocus on the leg. The spectrum was collected using the $\log\frac{1}{R}$ setting and was saved as a JDX file.

General Collection Procedure

All spectra were acquired at 20–23°C and at a humidity interval of 19.3–45.3%. A background spectrum was acquired before collecting any sample spectra. Sample spectra collection was then started. A new background spectrum was acquired every 20 samples or every 25 minutes, whichever condition was satisfied first, to account for changing variables in the environment. New back-

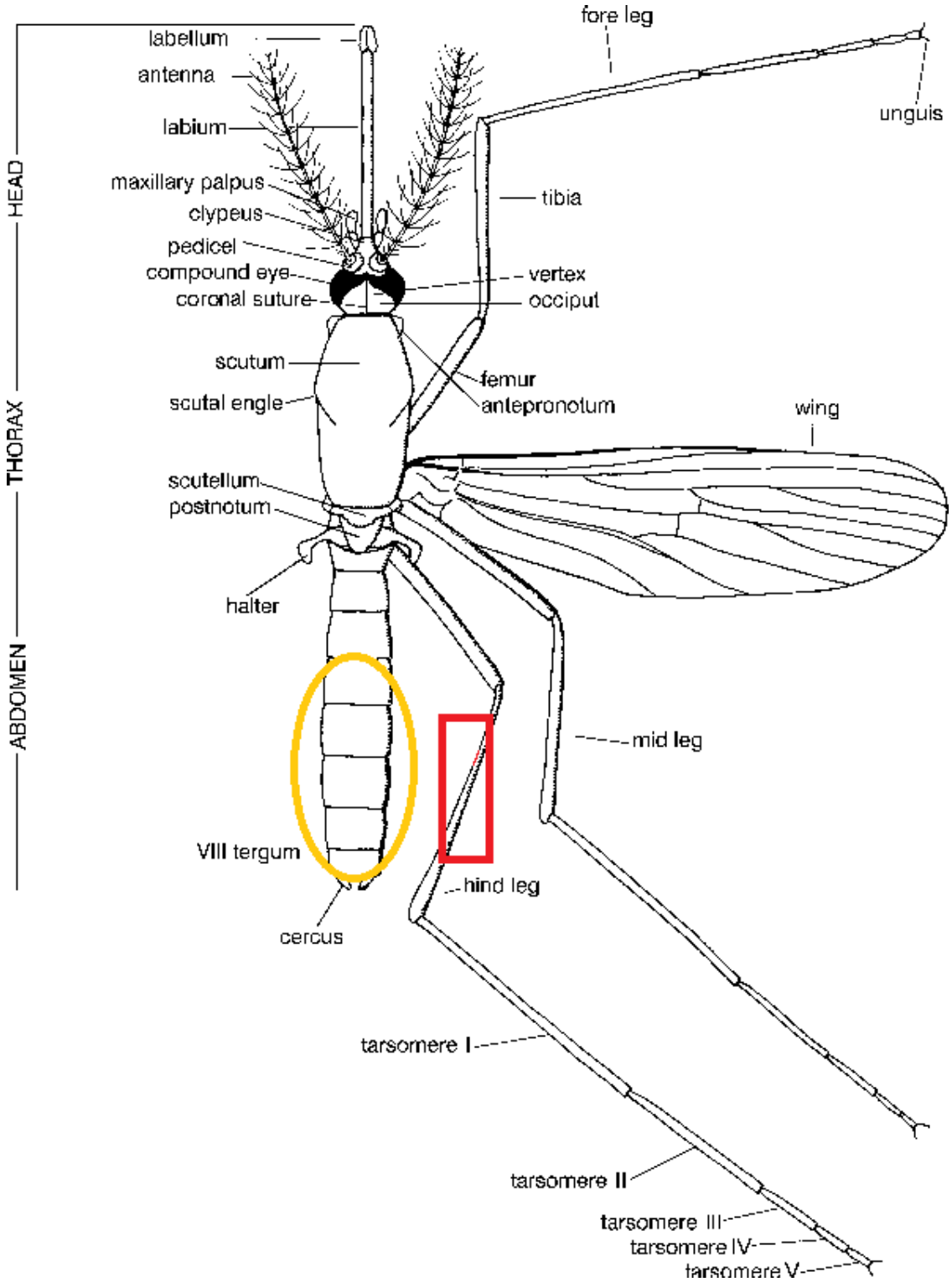

Figure 3. Probed region of mosquito anatomy by IR microspectroscopy represented by boxed in region, and region containing ovaries represented by circled region. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Morphology of Mosquitoes, Norbert Becker, Duan Petri, Marija Zgomba, Clive Boase, Minoo Madon, Christine Dahl, Achim Kaiser, 2010

ground spectra may have been collected before reaching 20 samples or 25 minutes when excessive water vapor interference was observed in the sample spectrum. If water vapor interference was observed, a new background spectrum was collected, then the affected sample spectrum was recollected. Each sample was measured, and the files were saved in the JDX file format. The metadata for the samples were recorded. The metadata contained information about each spectrum or the sample itself: sample identification, time between collection of background spectrum and sample spectrum, temperature, humidity, and optionally, miscellaneous notes. The miscellaneous notes were used to detail striking observations like unexpected spectra or poor conditions on a day of measurement. The JDX files containing the spectral data of the mosquitoes and the metadata were used to compile a Hierarchial Data Format[27] (HDF) file that would be loaded for further data processing.

Data Preprocessing Method

An HDF[27] file containing the spectral data of all mosquitoes and accompanying metadata was used for processing. The mosquito spectral data was preprocessed by means of outlier removal, cropping, normalization, and smoothing. Outlying spectra were removed visually after wavelength comparison. Additional outliers were removed from the first two days of data collection due to poor technique. Fifty spectra from the set of parous mosquito spectra were removed at random through a randomizer to even out the sample size imbalance. The spectral data was cropped 3100–650 cm^{-1}. Additional cropping regions with relevant chemical features were selected, such as peaks representing proteins and others that represented lipids. The spectra were normalized where the highest absorbance (the absorbance at 1653 cm^{-1}) was set to 1, and the lowest absorbance was set to 0. The SG algorithm was then applied to the spectral data using window size of 11, polynomial order of 2, and a derivative order of 2.

Data Processing Chemometric Methods

The preprocessed data were fit to PCA and PLS-DA models. Due to PCA performing poorly, the PCA data was used as an exploratory method prior to PLS-DA. The accuracy, true positive rate,

and true negative rate of the data derived from PLS-DA were recorded.

Additional Cropping and PLS-DA

The cropped data set was cropped further to create new spectral windows to test for PLS-DA based discrimination at different spectral regions while maintaining the same parameters for normalization and smoothing. The data was cropped deliberately to close in on spectral regions that attributed to best PLS-DA based discrimination.

CHAPTER THREE: RESULTS AND DISCUSSION

Raw Spectra

In Figure 4, all spectra of the mosquitoes plotted prior to processing are shown (170 spectra). To prepare the data for chemometric analysis, the spectra were preprocessed to standardize the data and eliminate outliers that could skew the distribution of the validation and training sets which would result in a poorer fit for the model. Preprocessing also de-emphasized spectral baseline fluctuations due to differences in the geometries of the samples. In Table 2, the peaks found in the spectra are assigned.

Outlier Identification, Outlier Removal, and Imbalance Correction

Outliers were first removed by visual inspection wherein spectra that obviously deviated from the average mosquito spectrum were dropped. More outliers were identified by their separation from the central data cluster in PCA score plots and removed. Although PCA was initially conceptualized to be used as a chemometric data analysis technique, PCA performed poorly in clustering parous and nulliparous mosquitoes separately. In Figure 5, a red ellipse was used to indicate the three outliers in a PCA scores plot of the spectra. The data from the first two days of collection, which consisted of 25 nulliparous mosquito spectra, were removed due to improper spectral collection technique. Fifty spectra of parous mosquitoes were removed at random to correct the imbalance in the datasets. After outlier removal and dataset imbalance correction, there were 46 parous and 41 nulliparous mosquitoes.

Cropping, Normalization, and Smoothing Results

The spectra were cropped 3100–650 cm^{-1}, normalized by maximum absorbance, and smoothed using SG smoothing. In Figure 6, a representative spectrum was used to demonstrate the effect of each preprocessing step on the spectra. A representative spectrum was used to demonstrate the effect of these preprocessing steps instead of the average spectrum because the average spectrum does not demonstrate some of the preprocessing steps well, especially smoothing since averaging

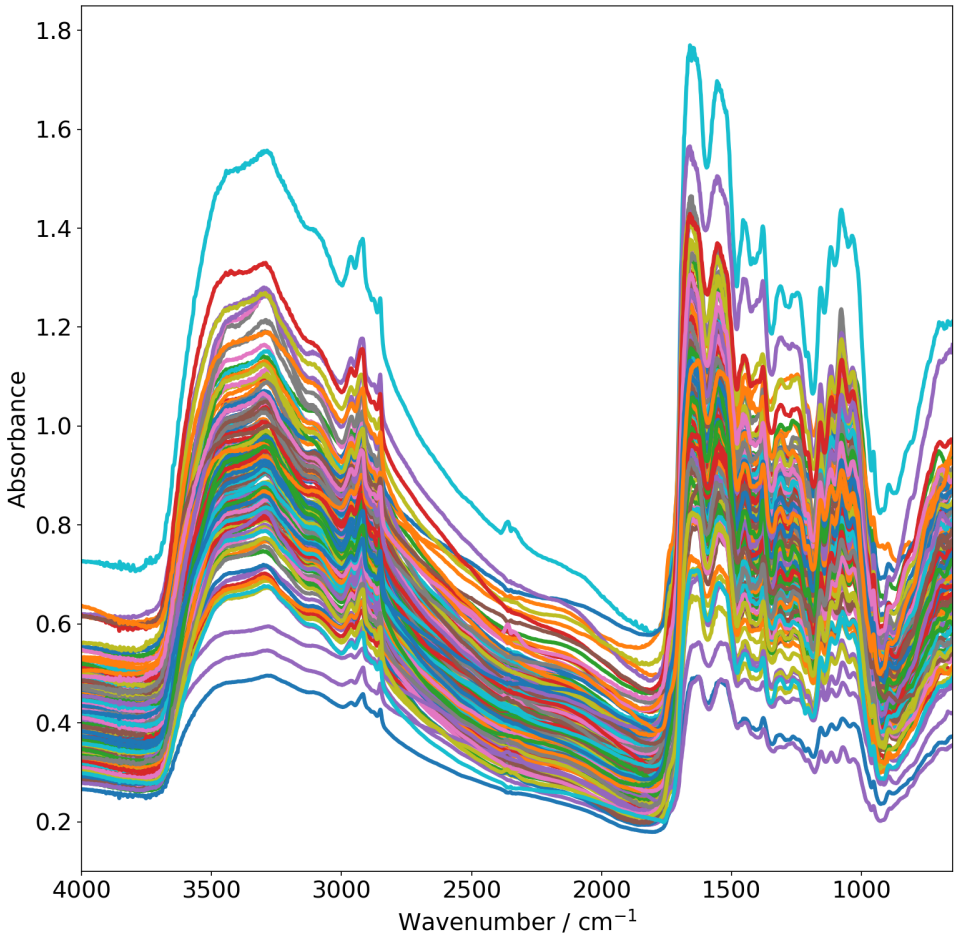

Figure 4. Spectra of mosquitoes prior to processing.

Table 2. Assignment of spectral peaks.

Wavenumber / cm^{-1}	**Assignment**	**Significance**	**Reference**
2962	CH_3 stretch	chitin	[28]
2920	C-H stretch (R_2CH_2)	lipid	[29]
	C-H stretch (RCH_3)	lipid	[29]
	C-H stretch	DNA	[30]
	$COCH_3$ stretch	chitin	[31]
2876	R_2-CH_2 (C-H stretch)	lipid	[29]
	R_3-CH (C-H stretch)	lipid	[29]
2892	CH_2 symmetric stretch of C-5	DNA	[30]
	C-H stretch	chitin	[28]
2852	C-H stretch (aldehyde)	lipid	[29]
1653	Amide I (C=O stretch)	protein	[32]
	alpha helix	protein	[33]
	C=C stretch	steroids	[34]
	C-O stretch	chitin	[31]
	C=O stretch of N-acetyl group	chitin	[31]
1550	amide II (NH bend)	protein, chitin	[31,33]
1451	C-H bend	lipid	[29]
1377	C-O stretch and C-OH bend	lipid	[29]
	CH bend, CH_3 bend	chitin	[28]
1302	C-O stretch and C-OH bend	lipid	[29]
1250	C-O stretch and C-OH bend	lipid	[29]
	CH_2 twist	DNA	[30]
1196	C-O stretch and C-OH bend	lipid	[29]
	CO stretch, CC stretch	DNA	[30]
	C-N assymmetric stretch (secondary alpha carbon)	protein	[32]
1157	C-O stretch and C-OH bend	lipid	[29]
	Bridge O asymmetric stretch	chitin	[28]
1118	C-O stretch and C-OH bend	lipid	[29]
		chitin	[28]
1076	C-O stretch and C-OH bend	lipid	[29]
		chitin	[28]
	C-N stretch	protein	[32]
1033	C-O stretch and C-OH bend	lipids	[29]
955	CH3 wag (along chain)	chitin	[28]
895	C-H bend	lipids	[29]
	CC stretch, CCH in-plane bend	DNA	[30]
700	NH bend	protein	[32,33]
	C-H bend	steroids	[34]
	N-H asymmetric bend	protein	[32]

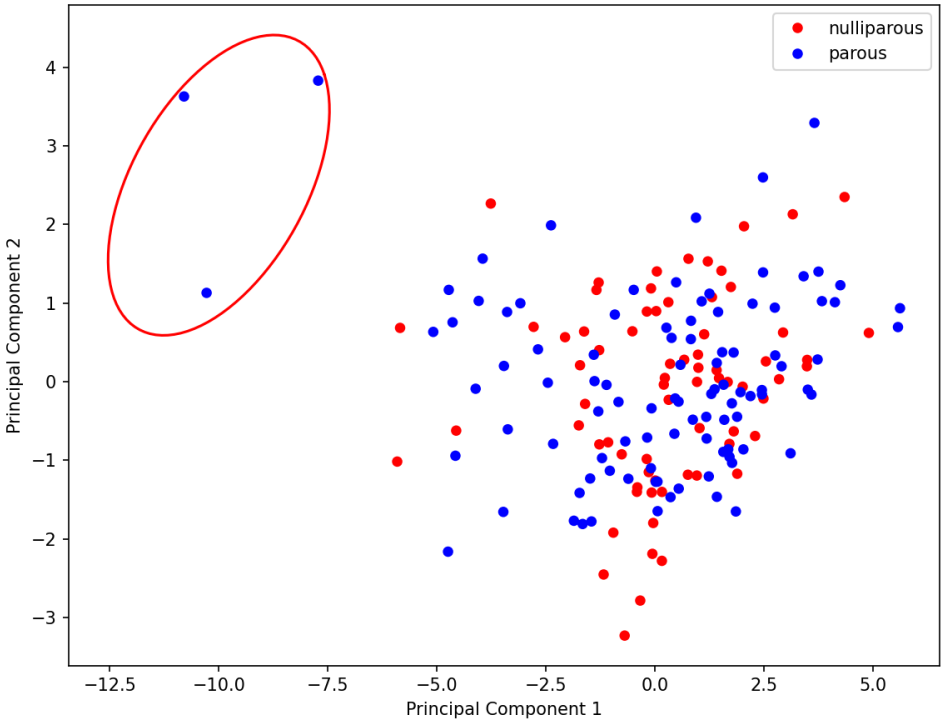

Figure 5. Scores plot of mosquitoes (170 samples) with outliers circled.

the spectra has similar effects to smoothing. In Figure 6a, the effect of normalizing the spectrum to the absorbance at 1653 cm^{-1} is shown (blue) compared to the spectrum before normalization (red). In Figure 6b, the effect of the normalized spectrum with an 11-point, second-order polynomial smoothing (orange) compared to the pre-smoothed spectrum (blue) is shown. By setting the derivative order to 0, the effect of only smoothing the spectrum without the added exaggerated effects of differentiating the data is shown. In Figure 6c, the effect of differentiating the smoothed spectrum to the 2nd order is shown. The differentiation of the smoothed data served as the final preprocessing step prior to PLS-DA. In Figure 7, the averaged spectra of parous and nulliparous spectra before preprocessing are shown, and in Figure 8, the averaged spectra of parous and nulliparous mosquitoes postprocessing, excluding SG, are shown.

Partial Least Squares Discriminant Analysis

Partial least squares discriminant analysis (PLS-DA) was used to classify and identify the dataset of parous and nulliparous mosquitoes. In Figure 9, the PLS-DA discrimination prediction is shown as a box plot. The vertical axis represents the actual identity of the set of mosquitoes where **np** is the set of nulliparous *Aedes triseriatus* mosquitoes and **p** is the set of parous *Aedes triseriatus* mosquitoes. The color of each marker on the plot represents the identity of each mosquito sample as determined by an entomologist (Brian Byrd, Western Carolina University) where red markers represent mosquitoes deemed to be nulliparous and blue markers represent parous mosquitoes. The data points were jittered so that each point can be seen. The dotted line (green) represents the threshold placed at 0.5. Because the data labels were input into the algorithm as **0** for nulliparous mosquitoes and as **1** for parous mosquitoes, 0.5 was chosen as the threshold as the natural midpoint between the two labels. Any point placed to the left of the threshold by the algorithm's prediction was counted as a predicted negative (predicted nulliparous), and those placed to the right of the threshold was counted as a predicted positive (predicted parous). The data was preprocessed by cropping 3100–650 cm^{-1}, normalization by tallest band, and second derivative SG.

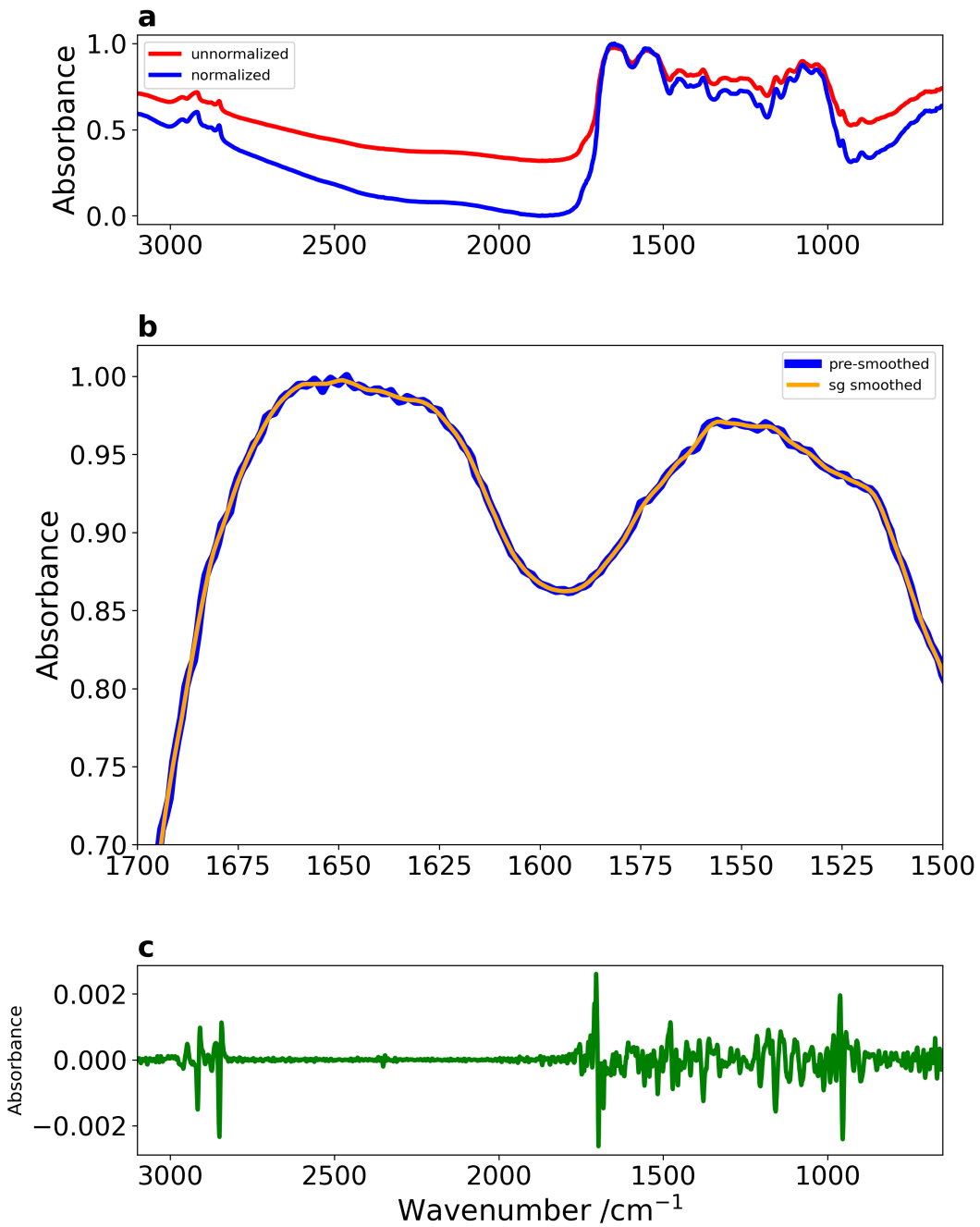

Figure 6. Spectra after each step of preprocessing on a representative spectrum.

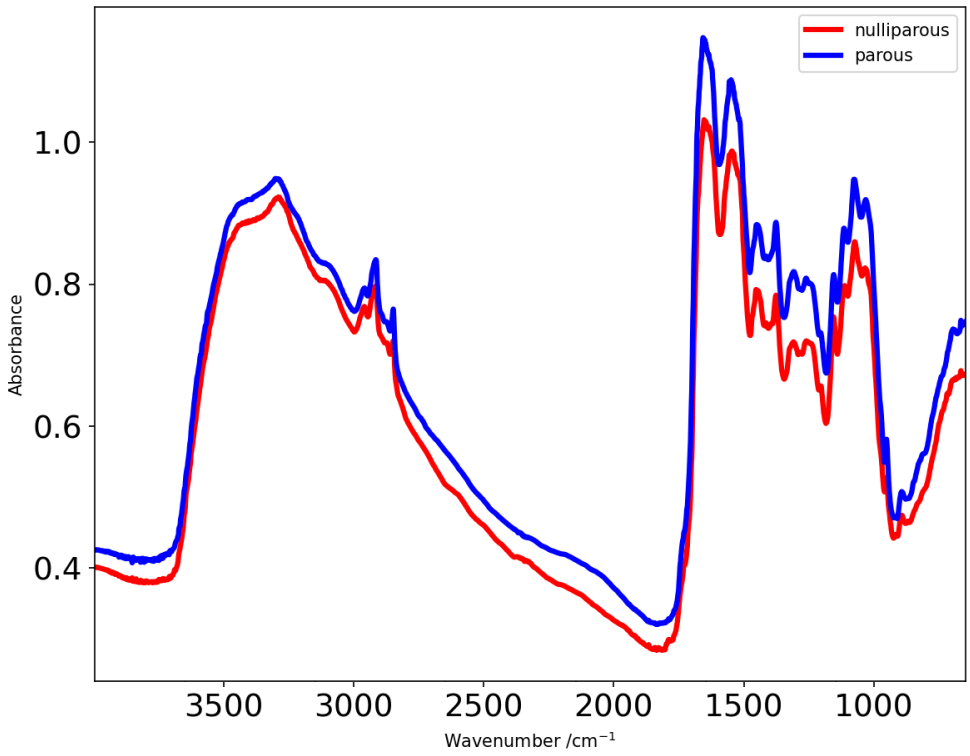

Figure 7. Before preprocessing average spectra of nulliparous mosquitoes (red) compared to parous mosquitoes (blue)

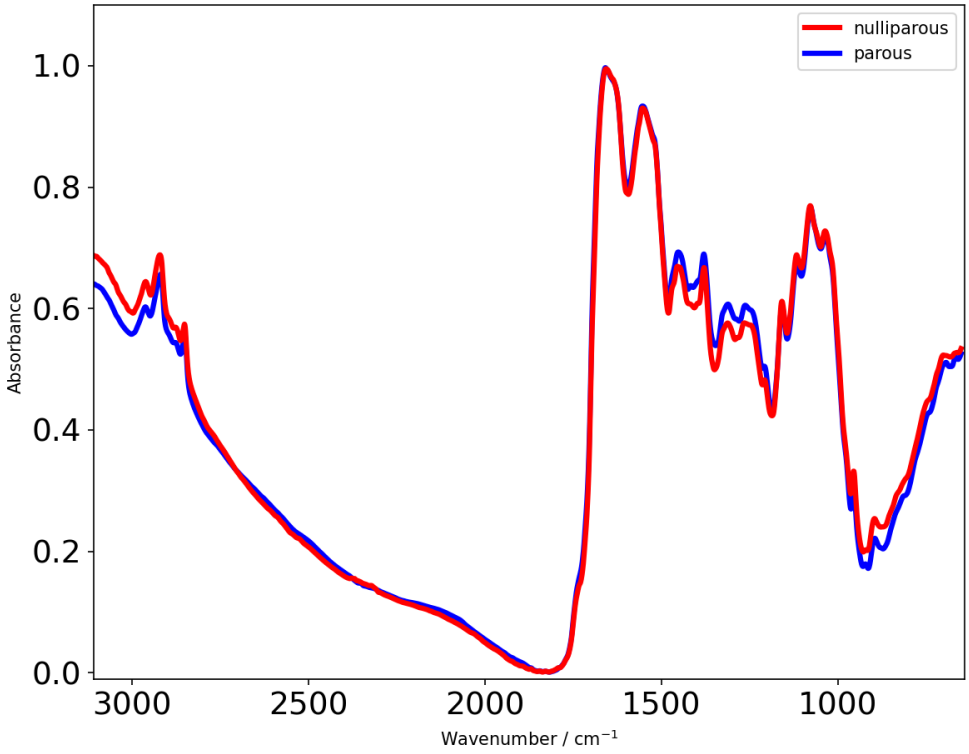

Figure 8. After preprocessing average spectra of nulliparous mosquitoes (red) compared to parous mosquitoes (blue).

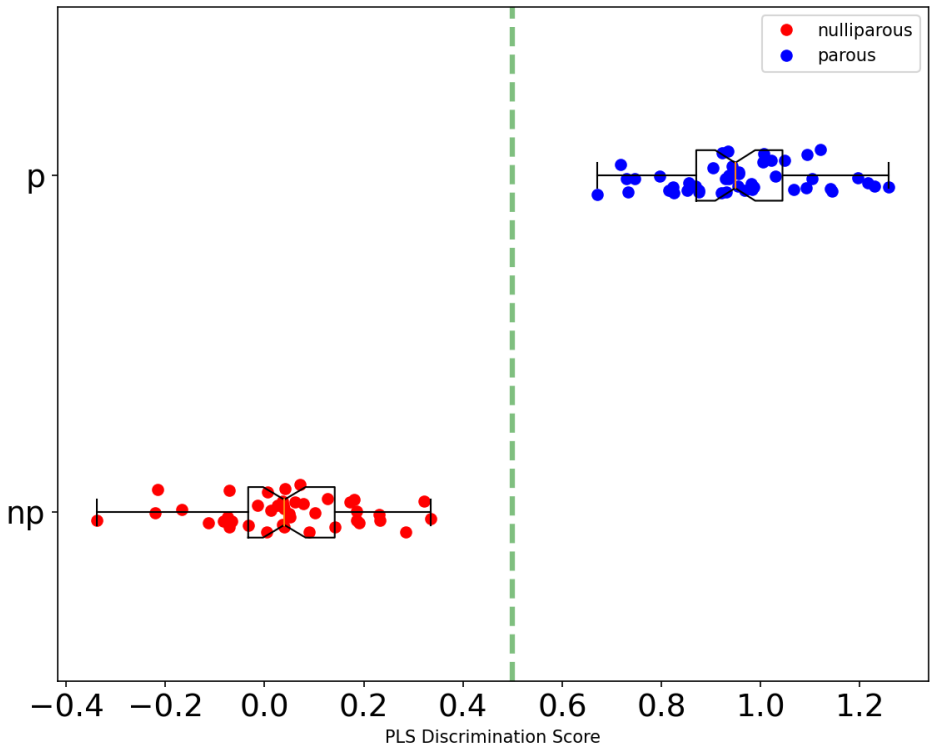

Figure 9. Box plot representation of PLS-DA data.

PLS-DA with Other Spectral Windows

Additional spectral regions were selected to test PLS-DA. Although the whole-spectrum window may suffice for successful discrimination, additional spectral windows can be used to determine what spectral regions are most useful in discrimination. In Table 3, the letters designating the spectral windows shown in Figure 10 are listed with their corresponding wavenumber interval and significance along with the PLS-DA accuracy, true positive rate (TPR), and true negative rate (TNR) of PLS-DA predictions. In Figure 11, box plot representations of three selected windows (B, C, and J) from Table 3 are shown.

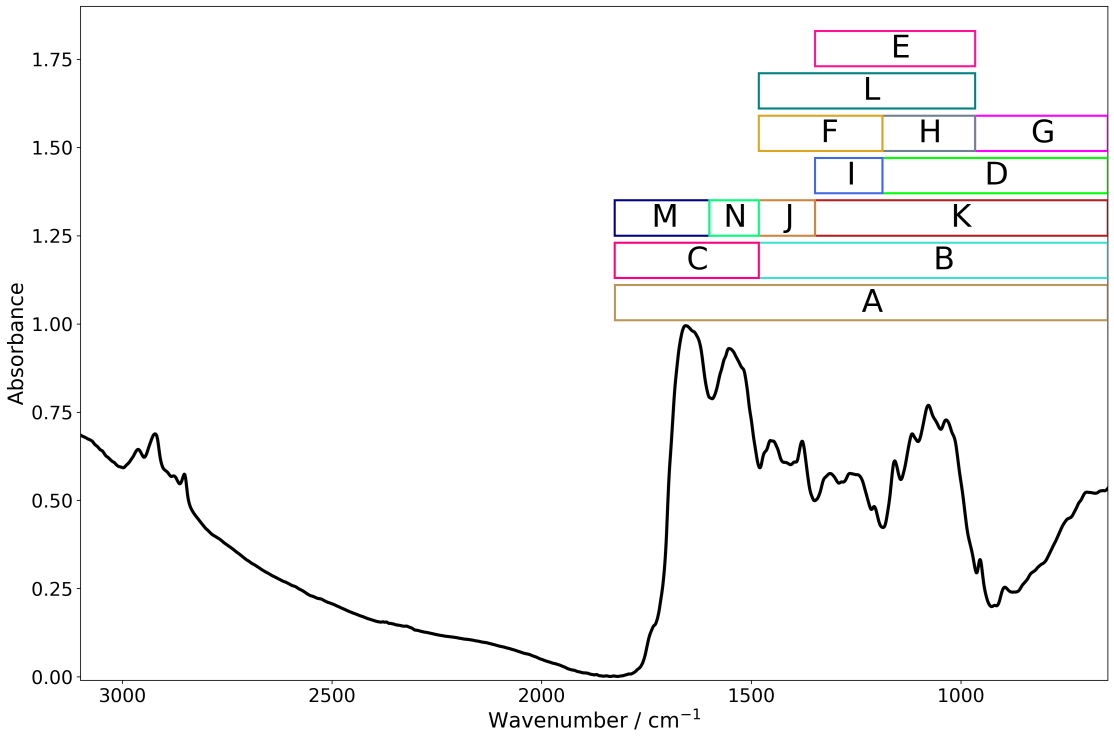

Figure 10. Average nulliparous spectrum with selected spectral windows.

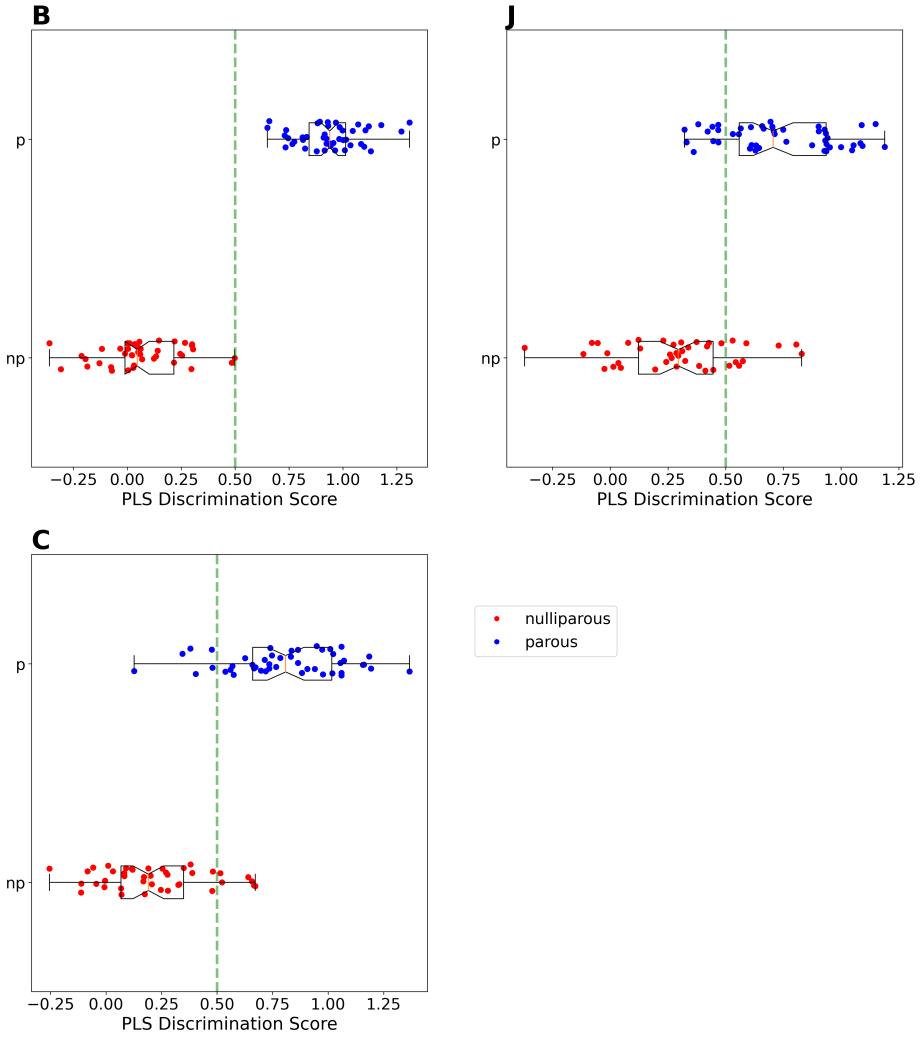

Figure 11. Box plot representation of PLS-DA results for three selected windows.

Table 3. Selected spectral regions PLS-DA results.

Window	Interval / cm^{-1}	Significance	Accuracy	TPR	TNR
–	3100–650	whole spectrum	100%	100%	100%
A	1826–650	chitin, DNA, lipids, protein, steroids	100%	100%	100%
B	1482–650	chitin, DNA, lipids, protein, steroids	100%	100%	100%
C	1826–1482	chitin, protein, steroids	86.2%	87.0%	85.4%
D	1187–650	chitin, DNA, lipids, protein, steroids	98.9%	100%	97.6%
E	1348–966	chitin, DNA, lipids, protein	97.7%	100%	95.1%
F	1482–1187	chitin, DNA, lipids, protein	95.4%	95.7%	95.1%
G	966–650	chitin, DNA, lipids, protein, steroids	92.0%	89.1%	95.1%
H	1187–966	chitin, lipids, protein	94.3%	97.8%	90.2%
I	1348–1187	DNA, lipids, protein	89.7%	89.1%	90.2%
J	1482–1348	chitin, lipids	78.2%	78.3%	78.0%
K	1348–650	chitin, DNA, lipids, protein, steroids	100%	100%	100%
L	1482–966	chitin, DNA, lipids, protein	95.4%	97.8%	92.7%
M	1826–1600	chitin, protein, steroids	82.8%	84.8%	80.5%
N	1600–1482	chitin, protein	86.2%	91.3%	80.5%

Discussion of Chemometric Analysis Results

When using the entire spectrum for PLS-DA discrimination, the results of the performance metrics were 100% accuracy, 100% TPR, and 100% TNR as shown in Figure 9 and Table 3. Further optimizations were made to test PLS-DA discrimination using the smallest possible spectral window while maintaining the performance from the whole-spectrum discrimination. The examples shown in Figure 11 were chosen to highlight the process by which the optimal windows were decided and to show how PLS-DA on spectral windows with different performance metrics appeared on box plots. Shown in Figure 11J is an example of a PLS-DA that performed relatively poorly as a means of contrasting with those that peformed well. As shown in Table 3, PLS-DA on window **A** returned the performance metrics of 100% accuracy, 100% TPR, and 100% TNR. Windows **B** and **C** were regions derived from window **A**. PLS-DA on window **B** had 100% accuracy, TPR, and TNR, whereas PLS-DA on window **C** had 86.2 % accuracy, 87.0% TPR, and

85.4% TNR. The reduction in accuracy, TPR, and TNR in window **C** as compared to windows **A** and **B** indicates that the optimal window is more likely to be contained in window **B**, so window **B** can be further reduced to find the smallest optimal window size. Windows **D–J** and window **L** were smaller windows derived from window **B**, but they did not maintain the performance metrics of window **B**. Window **K** had the smallest possible window size that maintained the 100% performance metrics. By finding the smallest possible window, insights into what chemicals are affected by mosquito parity can be made. These insights are limited, however, due to covariance in IR spectra where a compound may be represented on multiple areas of the spectrum, and several compounds may be represented on a single area of the spectrum. For example, window **K**, which was found to be the optimal window in this study, contains spectral features that can be attributed to lipids, DNA, protein, chitin, and steroids. Further study would be required to deliniate the role of these compounds in parity discrimination using PLS-DA.

PLS-DA Scores Median Analysis

An advantage of representing the PLS-DA scores on a box plot is facilitating statistical analyses at a glance. Shown in Figure 12 is a plot of the absolute error calculated using the median of the PLS-DA scores (which can be seen as the 2nd quartile in the box plots in Figure 9 and Figure 11) of parous mosquitoes versus TPR, and shown in Figure 13 is a plot of the absolute error calculated using the median of the PLS-DA scores of nulliparous mosquitoes versus TNR. Because it can be expected that the median of the PLS-DA scores data should fall near the designation determined in the class labels (**1** for parous and **0** for nulliparous), the absolute error can be thought of as an auxillary performance metric. Absolute error is calculated with the following general equation[35]

$$\text{Absolute error} = |\text{true value} - \text{measured value}| \tag{5}$$

which was adapted to the true value being the class label designation, **1** or **0**, and the measured value being the median of parous or nulliparous PLS-DA scores data respectively. TPR was used to compare to the absolute error in parous mosquitoes because TPR is a performance metric of the algorithm's ability to correctly identify parous mosquitoes, and similarly for TNR for the absolute error in nulliparous mosquitoes. The plots in Figure 12 and Figure 13 display expected correlation where higher TNR or TPR corresponds to lower absolute error, which gives additional confidence in the windows that were determined to be optimal.

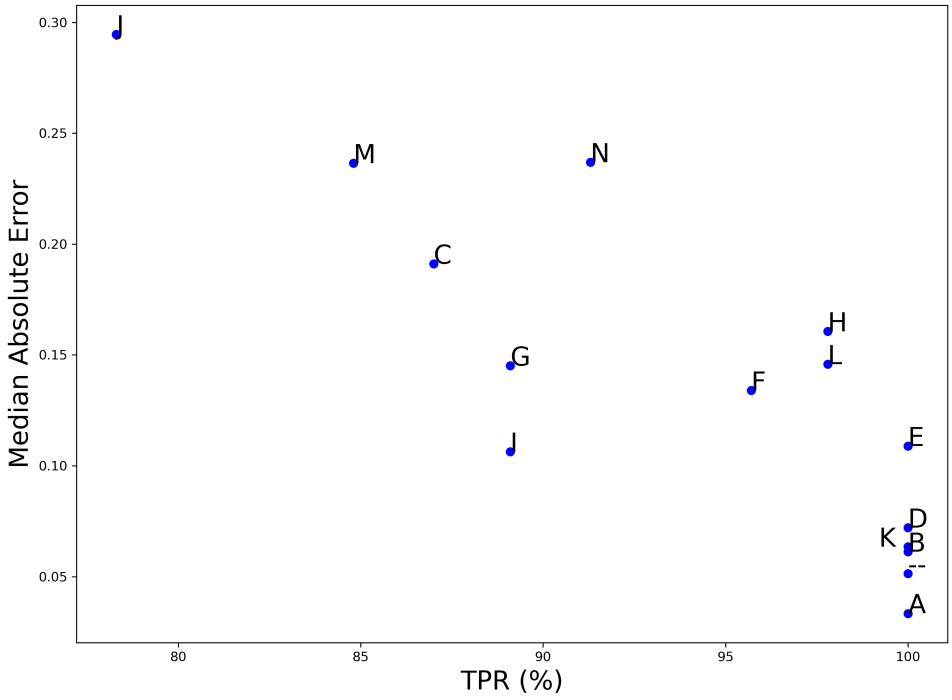

Figure 12. Parous statistics of PLS-DA showing absolute difference of the median vs. TPR and annotated with window designation.

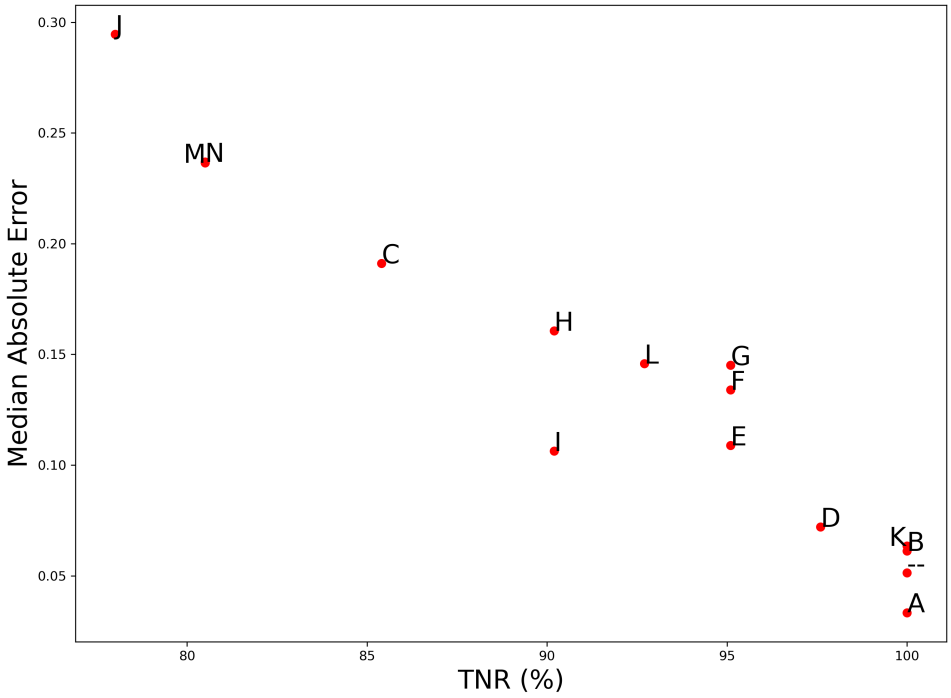

Figure 13. Nulliparous statistics of PLS-DA showing absolute difference of the median vs. TNR and annotated with window designation.

PLS-DA Scores Dispersion

Another statistical metric that can be analyzed from the box plots is dispersion in the form of interquartile range (IQR) which is calculated with the equation

$$IQR = Q3 - Q1 \tag{6}$$

where $Q3$ is the third quartile represented by the right-most edge of the box, and $Q1$ is the first quartile represented by the left-most edge of the box.[36] A table of each window and the IQR of the nulliparous and parous data are shown in Table 4. A window that has greater dispersion, for

Table 4. Selected spectral windows with IQR for parous and nulliparous data.

Window	nulliparous IQR	parous IQR
–	0.21	0.19
A	0.21	0.22
B	0.24	0.19
C	0.32	0.37
D	0.26	0.32
E	0.31	0.33
F	0.29	0.37
G	0.27	0.27
H	0.40	0.39
I	0.34	0.35
J	0.36	0.39
K	0.21	0.25
L	0.36	0.27
M	0.38	0.50
N	0.29	0.27

example window **M**, indicates that the algorithm detects greater differences between the samples of the same class designation compared to a window, like window **B**, with less dispersion. A

possible interpretation is that because the cropping of the region restricts the number of orignial variables that can be used in the creation of the latent variables, the dispersion may be greater if the chosen window does not contain as many of the original variables that are covariant between spectra of the same class designation. Although a high dispersion does not necessarily indicate a poorly performing model, further validation with more samples may be used to prove the performance of models that have good performance metrics but have high dispersion. Methods used for detecting statistically significant high dispersion are determined at the discretion of the researcher; however, in this study, none of the models with good performance metrics appeared to have worrisome dispersion warrenting such a test.

Possible Explanation of Poor Performance of PCA

The reason PCA performed poorly in this study compared to PLS-DA may be explained by a deeper look into the underlying algorithms governing their discrimination. PCA creates principal component variables that explain the greatest variation in the whole dataset without indication of the class designations. For the algorithm to correctly cluster parous mosquitoes as one cluster and nulliparous as another, the greatest variation in the dataset must be the latent variable "parity". However, if there is another variable (such as one of the natural variables that exist between individual samples) that is varied more in the samples than parity, the PCA algorithm will use that variable instead of parity to create the principal components. In contrast, PLS-DA creates latent variables that explain the greatest covariance between the training set and the class labels. The input of the class labels of parous versus nulliparous mosquitoes essentially gives the algorithm a "hint" that the latent variable that should be focused on is the parity of the mosquitoes instead of any other possible variation in the data. The supervision of the PLS-DA algorithm versus the unsupervised method of the PCA algorithm explains why PLS-DA performed better than PCA.

CHAPTER FOUR: CONCLUSION

This study shows the ability of PLS-DA to discriminate between parous and nulliparous *Aedes triseriatus* mosquitoes. The model was able to discriminate between parous and nulliparous *Aedes triseriatus* mosquitoes with 100% accuracy. The 100% accuracy performance metric of the PLS-DA model was maintained to a window as small as window **K** (1348–650 cm^{-1}), but further reducing of the window size resulted in poorer performance. This window, being the optimized window, contains all spectral signatures required for parity descrimination and has the same performance as using the entire spectrum. PCA did not work in this study as a discrimination method, but was used instead as an outlier detection tool. To improve this study, (1) the performance of the PLS-DA model can be validated with more samples that were measured using the optimal parameters (window **K**), (2) more samples can be included in the training set or used in a new training set to further optimize the model, (3) the method used in this study could be tested with wild-caught *Aedes triseriatus* mosquitoes, (4) other species of mosquitoes can be used to test this method, and (5) the loading vectors that provide information on how much each of the original wavelengths contributed most to the creation of the latent variables derived from the PLS-DA model could be used as a feature selection tool to improving cropping methods before repeating chemometric data analysis. Studies that can build upon this study could use this method to (1) discriminate parity in wild-caught *Aedes triseriatus* mosquitoes, (2) discriminate parity in other species of mosquitoes, (3) discriminate between information and statuses other than parity (such as infection status) in mosquitoes.

REFERENCES

[1] Rosenberg, R.; Lindsey, N. P.; Fischer, M.; Gregory, C. J.; Hinckley, A. F.; Mead, P. S.; Paz-Bailey, G.; Waterman, S. H.; Drexler, N. A.; Kersh, G. J.; Hooks, H.; Partridge, S. K.; Visser, S. N.; Beard, C. B.; Petersen, L. R. *Vital Signs*: Trends in Reported Vectorborne Disease Cases - United States and Territories, 2004 – 2016. *MMWR. Morb. Mortal. Wkly. Rep.* **2018**, *67*, 496–501.

[2] Franklinos, L. H. V.; Jones, K. E.; Redding, D. W.; Abubakar, I. The effect of global change on mosquito-borne disease. *Lancet Infect. Dis.* **2019**, *19*, e302–e312.

[3] Gubler, D. J. Resurgent vector-borne diseases as a global health problem. *Emerging Infect. Dis.* **1998**, *4*, 442–450.

[4] Sroute, L.; Byrd, B. D.; Huffman, S. W. Classification of Mosquitoes with Infrared Spectroscopy and Partial Least Squares-Discriminant Analysis. *Appl. Spectrosc.* **2020**, *74*, 900–912.

[5] Caputo, B.; Manica, M. Mosquito surveillance and disease outbreak risk models to inform mosquito-control operations in Europe. *Current Opinion in Insect Science* **2020**, *39*, 101–108.

[6] Milali, M. P.; Kiware, S. S.; Govella, N. J.; Okumu, F.; Bansal, N.; Bozdag, S.; Charlwood, J. D.; Maia, M. F.; Ogoma, S. B.; Dowell, F. E.; Corliss, G. F.; Sikulu-Lord, M. T.; Povinelli, R. J. An autoencoder and artificial neural network-based method to estimate parity status of wild mosquitoes from near-infrared spectra. *PloS one* **2020**, *15*, e0234557.

[7] Detinova, T. S. Age-grouping methods in Diptera of medical importance with special reference to some vectors of malaria. *Monograph series. World Health Organization* **1962**, *47*, 13–191.

[8] Charlwood, J. D.; Tomás, E. V. E.; Andegiorgish, A. K.; Mihreteab, S.; LeClair, C. 'We like it wet': a comparison between dissection techniques for the assessment of parity in Anopheles arabiensis and determination of sac stage in mosquitoes alive or dead on collection. *PeerJ* **2018**, *6*, e5155.

[9] Mollaoglu, A. D.; Ozyurt, I.; Severcan, F. In *Infrared Spectroscopy*; El-Azazy, M., Ed.; IntechOpen: Rijeka, 2019; Chapter 5.

[10] Lorenz-Fonfria, V. A. Infrared Difference Spectroscopy of Proteins: From Bands to Bonds. *Chem. Rev.* **2020**, *120*, 3466–3576.

[11] *Infrared Spectroscopy: Fundamentals and Applications*; John Wiley & Sons, Ltd, 2004; Chapter 1, pp 1–13.

[12] Johnson, J. B.; Naiker, M. Mid-infrared spectroscopy for entomological purposes: A review. *J. Asia-Pac. Entomol.* **2020**, *23*, 613–621.

[13] Stuart, B. H. In *Infrared Spectroscopy: Fundamentals and Applications*; Ando, D. J., Stuart, B. H., Eds.; John Wiley & Sons, Ltd, 2004; Chapter 2, pp 15–44.

[14] Lee, L. C.; Liong, C. Y.; Jemain, A. A. Effects of data pre-processing methods on classification of ATR-FTIR spectra of pen inks using partial least squares-discriminant analysis (PLS-DA). *Chemom. Intell. Lab. Syst.* **2018**, *182*, 90–100.

[15] Xu, S.; Lu, B.; Bell, N.; Nixon, M. Outlier Detection in Dynamic Systems with Multiple Operating Points and Application to Improve Industrial Flare Monitoring. *Processes* **2017**, *5*.

[16] Borille, B. T.; Marcelo, M. C. A.; Ortiz, R. S.; de Cssia Mariotti, K.; Ferro, M. F.; Limberger, R. P. Near infrared spectroscopy combined with chemometrics for growth stage clas-

sification of cannabis cultivated in a greenhouse from seized seeds. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2017**, *173*, 318–323.

[17] Delwiche, S. R.; Reeves, J. B., III. A Graphical Method to Evaluate Spectral Preprocessing in Multivariate Regression Calibrations: Example with Savitzky–Golay Filters and Partial Least Squares Regression. *Appl. Spectrosc.* **2010**, *64*, 73–82.

[18] Liberda, D.; Pita, E.; Pogoda, K.; Piergies, N.; Roman, M.; Koziol, P.; Wrobel, T. P.; Paluszkiewicz, C.; Kwiatek, W. M. The Impact of Preprocessing Methods for a Successful Prostate Cell Lines Discrimination Using Partial Least Squares Regression and Discriminant Analysis Based on Fourier Transform Infrared Imaging. *Cells* **2021**, *10*.

[19] Pierna, J. A. F.; Wahl, F.; de Noord, O. E.; Massart, D. L. Methods for outlier detection in prediction. *Chemom. Intell. Lab. Syst.* **2002**, *63*, 27–39.

[20] Saha, P.; Roy, N.; Mukherjee, D.; Sarkar, A. K. Application of Principal Component Analysis for Outlier Detection in Heterogeneous Traffic Data. *Procedia Computer Science* **2016**, *83*, 107–114.

[21] Bai, J.; Luo, J. SavitzkyGolay Smoothing and Differentiation Filter of Even Length: A Gram Polynomial Approach. *Spectroscopy* **2005**, *20*.

[22] Szymanska-Chargot, M.; Zdunek, A. Use of FT-IR Spectra and PCA to the Bulk Characterization of Cell Wall Residues of Fruits and Vegetables Along a Fraction Process. *Food Biophys.* **2013**, *8*, 29–42.

[23] Worley, B.; Powers, R. Multivariate Analysis in Metabolomics. *Curr. Metabolomics* **2013**, *1*, 92–107.

[24] Cordella, C. B. In *Analytical Chemistry*; Krull, I. S., Ed.; IntechOpen: Rijeka, 2012; Chapter 1.

[25] Tharwat, A. Classification assessment methods. *Appl. Comput. Inform.* **2021**, *17*, 168–192.

[26] Byrd, B. D. La Crosse Encephalitis: A Persistent Arboviral Threat in North Carolina. *North Carolina medical journal* **2016**, *77*, 330–333.

[27] The HDF Group, Hierarchical Data Format, version 5. 1997–2021; https://www.hdfgroup.org/HDF5/.

[28] Pearson, F. G.; Marchessault, R. H.; Liang, C. Y. Infrared spectra of crystalline polysaccharides. V. Chitin. *J. Polym. Sci.* **1960**, *43*, 101–116.

[29] Parker, F. S. *Applications of Infrared Spectroscopy in Biochemistry, Biology, and Medicine*; Springer US: Boston, MA, 1971; pp 142–164.

[30] Wiercigroch, E.; Szafraniec, E.; Czamara, K.; Pacia, M. Z.; Majzner, K.; Kochan, K.; Kaczor, A.; Baranska, M.; Malek, K. Raman and infrared spectroscopy of carbohydrates: A review. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2017**, *185*, 317–335.

[31] Majtán, J.; Bíliková, K.; Markovič, O.; Gróf, J.; Kogan, G.; Šimúth, J. Isolation and characterization of chitin from bumblebee (Bombus terrestris). *Int. J. Biol. Macromol.* **2007**, *40*, 237–241.

[32] Parker, F. S. *Applications of Infrared Spectroscopy in Biochemistry, Biology, and Medicine*; Springer US: Boston, MA, 1971; pp 165–172.

[33] Parker, F. S. *Applications of Infrared Spectroscopy in Biochemistry, Biology, and Medicine*; Springer US: Boston, MA, 1971; pp 188–231.

[34] Parker, F. S. *Applications of Infrared Spectroscopy in Biochemistry, Biology, and Medicine*; Springer US: Boston, MA, 1971; pp 315–349.

[35] Demmel, J. W. *Applied Numerical Linear Algebra*; Other Titles in Applied Mathematics; Society for Industrial and Applied Mathematics, 1997.

[36] Dekking, F. M.; Kraaikamp, C.; Lopuhaä, H. P.; Meester, L. E. *A Modern Introduction to Probability and Statistics: Understanding Why and How*; Springer Texts in Statistics; Springer London, 2006.