ABSTRACT

Chemometric Modeling of UV-Visible and LC-UV Data for Prediction of Hydrolysate Fermentability and Identification of Inhibitory Degradation Products

Negar Hedayatifar, Ph.D.

Mentor: C. Kevin Chambliss, Ph.D.

Production of ethanol from lignocellulosic biomass requires a pretreatment step to liberate fermentable sugars trapped within the plant. During pretreatment, lignin and some sugars undergo degradation to form compounds which have shown inhibitory effects to fermentative microorganisms. Accordingly, development of a rapid and accurate method for assessment of microbial inhibition and identification of inhibitory compounds is essential for gaining a better understanding of pretreatment and its downstream effects on fermentation processes.

Traditional methods for identification of inhibitory compounds involve a "bottom-up" approach. Using this approach, one or more known degradation compounds are added to fermentation media and their effects on batch fermentation of ethanol are observed. These methods are extremely time-consuming and labor-intensive which makes them unattractive to researchers. Furthermore, they are carried out on degradation compounds that have already been identified. Given that biomass hydrolysates contain many unidentified constituents, identification of inhibitory compounds by traditional means is unlikely to occur on a timescale that is consistent with current mandates for

commercial production of cellulosic ethanol. To address these limitations, we have developed a chemometric model that correlates ultraviolet (UV)-visible spectroscopic data of 21 different biomass hydrolysates with their fermentability (percent inhibition of ethanol production). This novel approach enables rapid prediction of hydrolysate fermentability using UV-visible spectroscopic data alone and offers significant improvements in throughput and labor when compared to traditional batch fermentation methods. The model was subsequently used to predict percent inhibition for five hydrolysate samples, with a root-mean-square error of prediction of 6%.

To evaluate the use of chemometric modeling for identification of inhibitory compounds in biomass hydrolysate, a second model was developed to correlate HPLC-UV chromatographic data of the 21 hydrolysates with their percent inhibition. Detection was monitored at four specific wavelengths identified by the UV-visible model as significant spectral regions. Once constructed, the HPLC-UV model was used to identify retention times that had the highest correlation with inhibition. To determine whether better resolution or more universal detectability of sample constituents may lead to identification of additional retention times, a third chemometric model was developed with chromatographic data of hydrolysates obtained via ion chromatography with conductivity detection.

Chemometric Modeling of UV-Visible and LC-UV Data for Prediction of Hydrolysate Fermentability and Identification of Inhibitory Degradation Products

by

Negar Hedayatifar, B.A.

A Dissertation

Approved by the Department of Chemistry and Biochemistry

_____

Patrick J. Farmer, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

_____

C. Kevin Chambliss, Ph.D., Chairperson

_____

Kenneth W. Busch, Ph.D.

_____

Stephen L. Gipson, Ph.D.

_____

Charles M. Garner, Ph.D.

_____

Sascha Usenko, Ph.D.

Accepted by the Graduate School
December 2011

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my husband Ali for his unwavering love and support during this endeavor. He has not only been a constant source of encouragement and strength in my life, but his positive attitude helped me see the world in a different view. Without him, this day would not have been possible.

I would also like to thank my advisor, Dr. Kevin Chambliss, for his supervision and support during my graduate studies at Baylor. I am especially grateful for his advice and encouragement during the writing of my dissertation. I have become a better writer because of him.

My sincere appreciation goes to Baylor University and the Department of Chemistry and Biochemistry for giving me the opportunity and financial support to study at Baylor. I am grateful to Dr. Christopher Becker for his guidance and compassion throughout my graduate studies at Baylor. I would also like to extend my gratitude to my committee members, Dr. Kenneth Busch, Dr. Stephen Gipson, Dr. Sascha Usenko, and Dr. Charles Garner for reading my dissertation and helping me improve it. I would especially like to thank Dr. Kenneth Busch for his patience, mentorship, and guidance over the years. My understanding of chemometrics would not be the same without him.

Last, but surely not least, I would like to thank my family for their encouragement throughout this journey. In particular, Sima, Setareh, and Jahan, for always seeing the best in me and cheering me on, not only in this journey, but also in life.

CHAPTER ONE

Hydrolysis and Fermentation of Lignocellulosic Materials for Ethanol Production

Ethanol as a Renewable Source of Energy

Crude oil has been the world's main energy source for decades. Given that energy consumption continues to increase, and that the rate of fossil fuel formation is significantly lower than the rate of expenditure, it can be expected that the world's petroleum reserves will be depleted at some point in the future.[1-2] Annual global oil production has been predicted to decline dramatically by year 2050.[3] Ethanol, on the other hand, is a renewable energy source produced via fermentation of sugars. It has been widely used in the United States as a fuel additive up to 10% by volume. Ethanol is renewable, burns cleaner than gasoline, and reduces the net carbon dioxide that is added to the atmosphere.[4-8] Due to its broad range of economical and environmental benefits, production of fuel ethanol from lignocellulosic biomass has been the subject of extensive research.[9,10] Lignocellulosic biomass is composed of lignin (15-20%), cellulose (40-50%), and hemicellulose (25-35%). Lignin is a cross-linked, hydrophobic macromolecule that contributes to the strength and rigidity of the plant cell wall. Cellulose and hemicellulose contain the fermentable sugars in oligomeric form, and represent the majority of a feedstock's composition. The abundance of fermentable sugars in lignocellulosic biomass makes it an attractive renewable resource for ethanol production. Conversion of biomass to ethanol is generally accomplished in three steps[11] (Figure 1.1): 1) pretreatment of biomass to liberate cellulose and break down

hemicellulose and lignin, 2) enzymatic hydrolysis of cellulose, and 3) fermentation of sugars to produce ethanol.

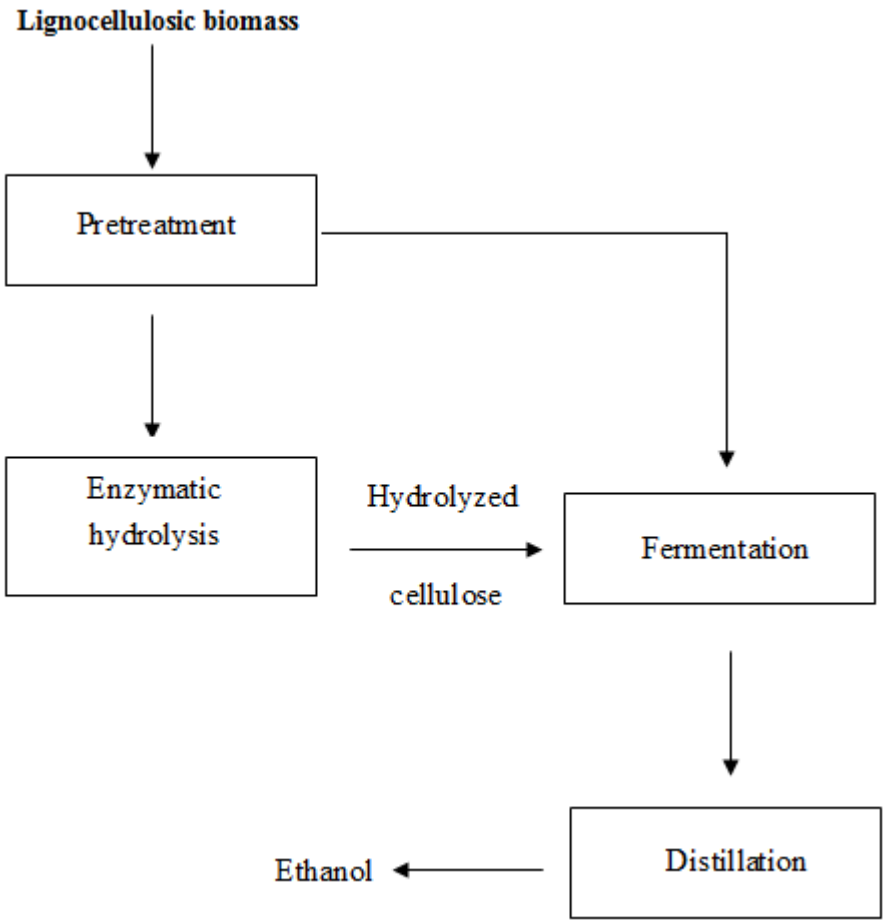


Figure 1.1. Typical scheme for production of ethanol from lignocellulosic biomass.

During pretreatment, lignin is broken down and hemicellulose is hydrolyzed to five and six-carbon sugars.[13] Cellulose, which is a polymer of β-1,4 linked glucose units, becomes accessible to further enzymatic hydrolysis. Hemicellulosic sugars are present in a soluble fraction called the hydrolysate, and cellulose remains in the insoluble fraction. Pretreatment is often carried out under harsh conditions, such as dilute sulfuric acid hydrolysis at elevated temperatures. Under these conditions, lignin and fermentable

sugars form degradation compounds that have shown inhibitory effects to downstream enzymatic and fermentation processes.[12-14] For example, glucose and xylose can undergo degradation to form 5-hydroxymethylfurfural (5-HMF) and furfural, respectively. Figure 1.2 shows an example of glucose degradation to 5-HMF.

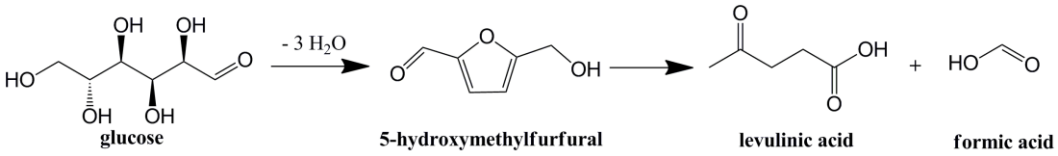


Figure 1.2. Formation of degradation products from glucose (reproduced from [15]).

The nature and concentration of degradation products formed largely depends on the feedstock type and the chemistry and severity of the pretreatment method employed.[13,16] A lack of a firm understanding regarding the identity of inhibitory compounds generated during pretreatment and their mechanism of action is one of the key barriers to cost-effective production of bioethanol.[13] Thus, identification of inhibitory hydrolysate constituents is a crucial step in gaining a better understanding of pretreatment and its downstream effects on enzymatic and fermentation processes. In this chapter, we provide a review of recent work related to identification of biomass degradation products and their inhibitory effects on ethanol production.

Biomass Composition

A wide variety of lignocellulosic feedstocks exist for production of bioethanol (e.g. hardwood, softwood, agricultural residues, and forestry residues). Lignocellulosic materials are composed of lignin, cellulose, and hemicellulose. Lignin is an aromatic polymer composed of condensed monomeric residues such as trans-*p*-coniferyl alcohol,

trans-*p*-coumaryl alcohol, and trans-*p*-sinapyl alcohol. The terms guaiacyl (G), *p*-hydroxyphenyl (H), and syringyl (S) are used to denote the three types of monomeric lignin residues.[13] Cellulose is composed of glucose units linked by β-1-4 glycosidic bonds.[17] Hemicellulose is a highly branched heteropolymer of pentose and hexose sugars. In addition to lignin, cellulose, and hemicellulose, plant biomass is also composed of extractive (water or ethanol-soluble components) and non-extractive (ash, pectin, proteins, and starch) components. Given the complex nature of biomass, obtaining a detailed compositional analysis is essential for evaluating its potential as a feedstock for production of bioethanol.

Compositional Analysis of Biomass Feedstock

Conventional wet chemistry methods for compositional analysis of biomass feedstocks are generally based on gravimetric, colorimetric, and chromatographic techniques. Additional analytical procedures have recently emerged for compositional analysis of biomass.[18-24] In a study by Hames and coworkers, a chemometric model was developed, using near infrared (NIR) spectroscopy and partial least squares (PLS) regression, for rapid prediction of 10 chemical properties (glucan, xylan, lignin, protein, acetyl, structural inorganics, galactan, arabinan, mannan, and uronic acids) for a wide range of corn stover samples.[18] The chemometric model demonstrated precision and accuracy similar to that of traditional wet chemistry methods. Pordesimo and coworkers further used this developed model to investigate correlations between corn stover composition and crop maturity.[19] Two identical corn cultivars were studied from two weeks before corn kernel physiological maturity until four weeks after the grain had reached the moisture content suitable for harvesting. Compositional analysis was carried

out using the previously established NIR model[18] and revealed a rapid decrease in soluble solids and increase in xylan and lignin content, immediately following crop physiological maturity.

Characterization of water-soluble materials in biomass feedstocks has also been reported in recent studies.[20,21] In a study by Chen and coworkers, over 30 previously unknown water-soluble constituents of corn stover were successfully characterized.[20] Five representative corn stover samples of varying composition were obtained, extracted (Soxhlet method), and analyzed using multiple analytical protocols. Overall, mass percent of water-soluble materials ranged from 14 to 27 % of the dry weight of the feedstock, with carbohydrates representing the largest fraction of the dry weight of extractives (30-57%). Alditols, aromatic acids, aliphatic acids, and inorganic cations and anions accounted for 3-7%, 0.06-0.1%, 7-21%, 6-14%, and 2-5% of the dry weight of extractives, respectively. Additionally, a band of material, with a reddish-brown color was consistently retained on the SPE cartridge, and was referred to as the red-brown fraction. This fraction was eluted with a 50:50 water-acetonitrile mixture, followed by gravimetric analysis of the residue remaining after solvent evaporation. The red brown fraction accounted for 10-18% of the dry weight of extractives. Qualitative analysis of this fraction via RPLC-UV suggested the presence of a complex oligomeric mixture. The authors suggested that the constituents of the red-brown fraction should be categorized in future compositional studies of biomass. In another study by Chen and coworkers, composition of water-soluble extractives in four switchgrass samples were analyzed.[21] An accelerated solvent extraction (ASE) system was used for extraction of water-soluble materials. This system demonstrated similar removal of water-soluble extractives and a

7-fold reduction in time compared to the Soxhlet technique. Overall, most compound classes exhibited similar contributions to the mass balance for water extractives compared to that observed for corn stover,[20] with carbohydrates and the red-brown fraction being the major exceptions. Monomeric sugars and the red-brown fraction accounted for 18-27% and 30-35% of the dry weight of extractives in switchgrass, respectively, compared to 30-57% and 10-18% in corn stover.[20] Overall, water-soluble extractives accounted for 12-15% of the dry weight of the feedstock.

In a recent study, rapid determination of lignin content using chemometric modeling of Fourier transform mid-infrared (FTIR) data was demonstrated by Tamaki and Mazza.[23] Total lignin content was defined as the amount of acid-insoluble and acid-soluble lignin. Acid-insoluble lignin was determined gravimetrically and acid-soluble lignin was determined using absorbance of hydrolysate at 320 nm. Two chemometric models were developed correlating the FTIR spectra of 67 triticale and 47 wheat straw samples with their lignin content (% wt/wt, oven-dry basis). A third model was constructed using all triticale and wheat samples. For all three models, 75% of samples were selected for calibration, and the remaining 25% were used to validate the models. All models were constructed by regressing FTIR data with lignin content, using partial least squares (PLS) algorithm. The predictive performance of the models was evaluated using the validation samples. Root-mean-square error of prediction (RMSEP) for the triticale, wheat, and combined model were 0.305, 0.163, and 0.272%, respectively, demonstrating that mid-infrared spectroscopic techniques can be used for rapid prediction of lignin content in triticale and wheat straw feedstocks.

In 2011, DeMartini and coworkers reported the development of a downscaled method for compositional analysis of biomass.[24] The method is based on the conventional wet chemistry methods developed by the National Renewable Energy Laboratory,[25] but is scaled down by a factor of 100 to use significantly less materials. The procedure is performed in HPLC vials and can be automated, thus, improving labor input and throughput. The major changes in the downscaled approach are: 1) centrifugation and decanting to accomplish solid-liquid separation as opposed to filtration, 2) omission of mixing in the first hydrolysis step, and 3) measurement of total acid-insoluble residue (lignin and ash). Comparison of the downscaled approach with traditional methods revealed identical carbohydrate composition and similar estimates of lignin and ash content for three biomass samples (poplar, switchgrass, and sugarcane bagasse).

Analysis of Degradation Compounds

Common degradation compounds produced during pretreatment of biomass include aliphatic acids, aromatic acids, aldehydes, and ketones.[26] Given their commercial importance, both as value-added products and potential fermentation inhibitors, a detailed understanding of formation and quantity of degradation products in biomass hydrolysate is critical. Several analytical methods have been developed for identification and quantitation of constituents in biomass hydrolysate. Gas chromatography techniques with flame ionization (FID) and mass spectrometry (MS) detection have traditionally been used for analysis of degradation products.[27-29] These methods require derivitization of target analytes prior to analysis, which is time-consuming and unattractive to researchers analyzing samples of unknown composition. Several high performance

liquid chromatography techniques (HPLC) are also available for analysis of degradation compounds in biomass hydrolysate, with UV, refractive index (RI), mass spectrometric, or conductivity detection, depending on the compound class being analyzed. Methods for simultaneous detection of degradation compounds from different analyte classes have also been developed. In a study by Chen and coworkers, an analytical procedure was developed, using methyl *tertiary*-butyl ether (MTBE) extraction of corn stover hydrolysate followed by HPLC analysis and UV detection.[30] The method allowed for simultaneous identification of 32 degradation compounds in a reference standard, and 15 compounds in a corn stover hydrolysate in a 120-minute run. In a more recent study, Sharma and coworkers expanded upon the method described above, using LC-photodiode array (PDA) and tandem mass spectrometry detection (LC-PDA-MS/MS) for simultaneous identification and quantitation of 40 degradation compounds in a 60-minute run.[31] In addition, Scarlata and Hyman have developed a fast (10-minute) HPLC method for the analysis of acetic acid, furfural, HMF, ethanol, and carbohydrates, using RI detection.[32]

It is important to note that the amount and nature of degradation products formed varies with each hydrolysate. The effects of varying pretreatment and feedstock type on formation of degradation products in biomass hydrolysate have been investigated recently.[16] Three feedstocks (corn stover, poplar, and pine) were pretreated using eight different methods (0.7% H_2SO_4, 0.07% H_2SO_4, liquid hot water, neutral buffer solution, aqueous ammonia, lime, lime with oxygen pressurization, and wet oxidation). Forty degradation products in the pretreated hydrolysates were analyzed using a previously established method by Sharma and coworkers.[31] The results demonstrated that the

formation of degradation compounds was a function of feedstock type, pretreatment chemistry and severity (pH, time, and temperature).

Assessment of Toxicity

Batch fermentation methods are typically employed for assessment of the toxicity of degradation products. Toxicity tests are generally carried out using a bottom-up experimental design, which involves the addition of one or more compounds to fermentation media and observing their effects on cell growth and ethanol production relative to a control. Inhibitory effects of several degradation compounds have been investigated previously.[13] Furanic compounds (i.e. furfural and 5-HMF), formed from degradation of sugars, have been of particular importance for toxicity in various microorganisms.[33-35] Furans inhibit glycolysis by interfering with three glycolytic enzymes, alcohol dehydrogenase (ADH), pyruvate dehydrogenase (PDH), and aldehyde dehydrogenase (AlDH). Furfural and HMF are metabolized by ADH to their corresponding alcohols, which results in a lag phase in cell growth.[36] Furfuryl alcohol only slightly inhibits anaerobic growth of *S. cerevisiae* and *Escherichia coli*,[37,38] but complete inhibition of aerobic growth of *Pichia stipitis* by furfuryl alcohol has been reported.[39] Furfural has been shown to inhibit ethanol production by 31% at concentrations as low as 1 g/L, when tested alone.[40] At 2 g/L, furfural further inhibited cell growth of *S. cerevisiae*. Inhibitory effects of HMF are less severe than furfural.[41] When tested alone, HMF has been shown to inhibit ethanol production by 20% at 2 g/L. When furfural and HMF were tested together, a total furan concentration of 3 g/L was required to inhibit ethanol production.[42] However, the minimum inhibitory concentration of furans is significantly lower when they are present in a mixture of other degradation

products, which may be due to synergistic effects. For example, Martinez and coworkers demonstrated that when the ratio of HMF to furfural is low (0.15-0.2), inhibition of ethanol production occurs at total furan concentration above 0.2 g/L.[42] Even as the ratio of HMF to furfural increased to 0.8-1.1, the minimum inhibitory concentration of furans was as low as 0.5 g/L. Several other studies have also demonstrated the inhibitory effects of HMF and furfural on cell growth and ethanol production in both pentose and hexose fermenting microorganisms.[43,44]

Acetic acid is another degradation compound that has been of particular importance in toxicity studies. Previous studies have demonstrated that acetic acid is inhibitory to ethanol production and cell growth.[28,34,45] Inhibitory effects of acetic acid are pH-dependent, since it is the undissociated form that can diffuse through the plasma membrane and dissociate in the cytosol.[46] Once dissociated, protons must be pumped across the membrane to maintain intracellular pH. This process requires energy, in the form of adenosine triphosphate (ATP). At low concentrations, acetic acid can increase the rate of ethanol production by increasing energy expenditure, and thus, sugar uptake. However, above a critical concentration of acetic acid, the diffusion rate of the undissociated form will exceed the proton pumping capacity of the plasma membrane, leading to cell death. Thus, increasing media pH can reduce the inhibitory effect of acetic acid.

In a study by Graves and coworkers, the effect of pH and lactic or acetic acid on ethanol production was investigated.[47] Corn mashes were supplemented with lactic or acetic acid at concentrations of 0, 0.5, 1.0, 2.0, 3.0, and 4.0 % w/v and 0, 0.1, 0.2, 0.4, 0.8, and 1.6% w/v, respectively. The pH values of all samples were subsequently

adjusted to 4.0 4.5, 5.0, or 5.5. Total ethanol concentration decreased in presence of 0.1% w/v acetic acid at pH 4. Complete inhibition of ethanol production by *S. cerevisiae* was observed at pH 4 and acetic acid concentration of 0.8% w/v or higher. Lactic acid inhibited ethanol production at a concentration of 3% w/v and pH ≤5. At 4% w/v, lactic acid inhibited ethanol production at all pH values.

Alkali and heavy metal salts are also present in lignocellulosic hydrolysates.[13] Potassium ion has been shown to inhibit hexokinase, enolase, and membrane ATPase activities and nucleic acid uptake in yeast cells.[49-51] Difference in toxicity for calcium, magnesium, sodium, ammonium, iron, cobalt, copper, zinc, sulfate, and phosphate ions have also been previously demonstrated.[13,52,53] Chloride ion has also been shown to have detrimental effects when present at concentrations exceeding 6 g/L.[48] In a recent study by Casey and coworkers, it was demonstrated that xylose fermentation is more sensitive to salt concentration than glucose fermentation, and that overall ethanol yield decreases with increasing salt concentration, due to the increase in formation of glycerol byproduct.[54] Although some ions demonstrate nutritive benefits to yeast metabolics and physiological functions, general consensus is that salts have a negative impact on microbial processes at elevated concentrations.

Synergistic Inhibition

Synergistic effects occur when the combination of two or more degradation products cause more toxic effects than the sum of their individual effects.[13] Several studies have investigated synergistic inhibition by testing different combinations of compounds and observing the inhibitory effects on cell growth and ethanol production.[13,26,37,55] Lactic acid has shown synergistic inhibitory effects when tested

together with acetic acid, especially at elevated temperatures (≥ 30°C).[56] Furfural has shown synergistic inhibition in *P. stipitis* and *S. cerevisiae*, when present in combination with acetic acid and HMF.[34,55] Furfural has also shown severely toxic effects on ethanol production by *E. coli* when present in combination with other aldehyde degradation products.[44] In addition, combinations of acetic acid with ethanol, and furfural with ethanol have demonstrated synergistic inhibition to ethanol producing microorganisms.[57] In a study by Oliva and coworkers, inhibitory effects of binary combinations of catechol, 4-hydroxybenzaldehyde, vanillin, and furfural on yeast *Kluyveromyces marxianus* were investigated.[58] Combinations of furfural with catechol and 4-hydroxybenzaldehyde demonstrated additive effects, while other binary combinations (catechol with 4-hydroxybenzaldehyde, and vanillin with catechol, furfural, or 4-hydroxybenzaldehyde) showed synergistic effects and caused a 60-90% decrease in cell growth. Vanillin demonstrated the most severe synergistic toxicity in binary combinations (90% reduction in cell growth and ethanol production). In another study by Oliva and coworkers, the synergistic effects of combinations of acetic acid, furfural and catechol on fermentation were investigated.[59] All compounds exhibited synergistic effects on ethanol production. Various combinations of sodium acetate, furfural, HMF, vanillin, and syringaldehyde have also exhibited synergistic inhibition on growth and H_2 production by *T. thermosaccharolyticum.*[60] It is important to note that most studies that investigated synergistic inhibition, utilized synthetic media for observing the effects of compounds studied, and the results are not always in agreement with synergistic inhibition that is observed in real hydrolysates.[61] This suggests that other degradation compounds, in addition to the few tested, may significantly contribute to microbial inhibition, and that a

more holistic approach is needed to identify the inhibitory constituents of biomass hydrolysate.

Detoxification Methods

Detoxification refers to the removal of inhibitors from the hydrolysate prior to fermentation. Several biological, chemical, and physical methods have been reported for detoxification of lignocellulosic hydrolysates. Several comprehensive reviews have been given previously.[13,26,62-64] Biological detoxification refers to the treatment of hydrolysate with enzymes, such as peroxidase and laccase, to increase ethanol productivity in hydrolysate.[65] Mechanism of detoxification by enzymes is suggested to be via oxidative polymerization of low-molecular-weight phenolic compounds. Certain bacterial and yeast strains are also capable of metabolizing fufural, HMF, and ferulic acid.[66] In 2005, Nichols and coworkers carried out a study investigating the ability of *Coniochaeta ligniaria* NRRL-30616 and 23 related fungal strains to metabolize furans and grow in dilute-acid hydrolysate of corn stover.[67] Seven of the 23 strains demonstrated the ability to grow using furfural as the main carbon source. *C. ligniaria* removed 99%, 85%, and 23% of furfural, HMF and acetate, respectively. However, 46%, 19%, and 15% of the glucose, xylose, and arabinose were also depleted. Subsequent fermentation of abated hydrolysate showed significant improvement over fermentation of unabated hydrolysate. In a recent study by Nichols and coworkers, the ability of *C. ligniaria* NRRL30616 to remove furans, organic acids, aldehydes, and phenolic compounds in corn stover hydrolysate was investigated.[68] Compounds from each of these classes were removed during the course of bioabatement. Furthermore, xylose metabolism during fermentation by *E. coli* was improved. Other detoxification methods, such as treatment with activated

charcoal or anion-exchange resins also effectively remove inhibitors.[62,69,70] A disadvantage of anion-exchange resins is that they require the pH of the hydrolysate to be adjusted to 10, which requires additional chemicals. Furthermore, use of anion-exchange resins can result in up to 26% loss of fermentable sugars. Although treatment with activated charcoal does not have the pH limitation, its high cost presents a disadvantage.

Alkali treatment has also been widely used for detoxification of hydrolysates. During this process, the pH of the hydrolysate is raised to 9-10 with $Ca(OH)_2$ (overliming) and subsequently adjusted to 5.5 with H_2SO_4.[13] In a study by Hames and coworkers, the use of ^{13}C-NMR spectroscopy for comparison of overlimed and nonoverlimed hydrolysates for identification of functional groups involved in the overliming reaction was investigated.[71] The spectra revealed that the major functional groups removed during overliming were lignin-derived aliphatic and aromatic acids and esters. Ketone and aldehyde functional groups were not detected in spectra of untreated hydrolysates. While overliming is effective at lowering toxicity of the hydrolysate, it results in the formation of an insoluble precipitate when used to detoxify dilute-acid (H_2SO_4) hydrolysate. Removal of this precipitate is carried out via centrifugation prior to fermentation, adding an additional (and costly) step to the overall process.[26] Overliming also leads to losses in sugar concentration. A review of several alkaline treatment studies, using various bases and process conditions, has been given recently.[26]

Multivariate Data Analysis

Most aspects of nature are multivariate. For example, a person's health is dependent on multiple factors, such as genetic predisposition, diet, exercise habits, stress level, etc. In chemical analyses, a single property of interest most often depends on

multiple factors. As such, it is needed to measure multiple variables at the same time. Multivariate data often contain information regarding the property of interest, referred to as "signal", and information that is not relevant, referred to as "noise".[72] The signal is not always apparent and may be hidden by the noise. Furthermore, collinearity in the data can be an issue, where the measured variables are intercorrelated and linearly dependent to some degree. Figure 1.3 shows an example of collinearity, where the responses at two different variables (wavelengths) for five different samples exhibit a linear trend. Ideally, for each variable to describe a unique set of information regarding the property of interest, the response at each measured variable would be independent of others. However, this is often not the case. Chemometrics can be a powerful tool for analyzing and extracting useful information from chemical data using statistics and mathematics.[73,74] Most common types of chemometric analyses include description, pattern recognition and classification, and regression and prediction.[75] The specific chemometric method employed will depend on the type of analysis required.

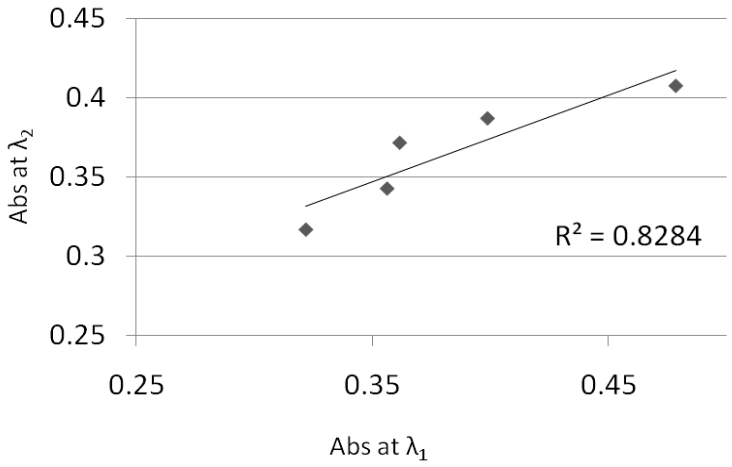


Figure 1.3. Hypothetical example showing collinearity in measured variables.

Multiple linear regression (MLR) is used to build chemometric models based on linear correlations between the measured data (independent variables) and the property of interest (dependent variable) (Figure 1.4).[72] In this way, every independent variable is associated with the dependent variable. The model is constructed using linear least squares and can be used to predict the value of a dependent variable when only the independent variables are provided.

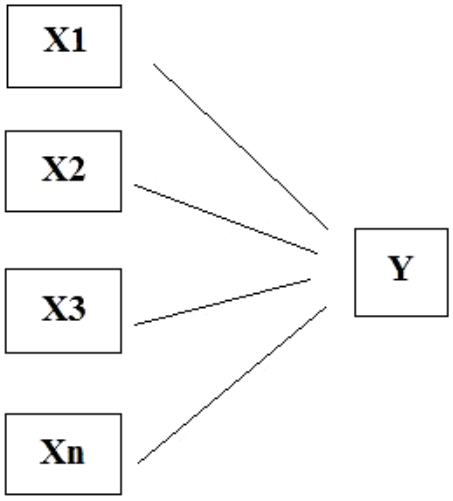


Figure 1.4. Hypothetical example of multiple linear regression analysis (reproduced from [73]).

Since MLR directly correlates each independent (X) variable with the dependent (Y) variable, any collinearity in the X data can have a detrimental effect on the accuracy of the model.[76] To avoid problems with collinearity, multivariate regression models should be constructed using an orthogonal coordinate system.[77] This can be accomplished using different algorithms such as principal component analysis (PCA), principal component regression (PCR), or partial least squares (PLS). PCA uses a few linear combinations of the independent variables that can be used to summarize the data

and explain the variance (measure of the spread of variables).[77] The linear combinations of the independent variables are used as axes to create a new coordinate system, thereby reducing the dimensionality of the data set. The new axes are referred to as principal components (PCs). The first PC is drawn in the direction of maximum variance through the independent data set, with each subsequent PC drawn in the next direction of highest variance and orthogonal to the preceding PC (Figure 1.5). Therefore, all PCs are orthogonal and have a common origin, but will contain different degrees of variance. The orthogonality constraint exists so that all PCs are independent of each other and each one explains a unique set of information, thereby eliminating the issue of collinearity.

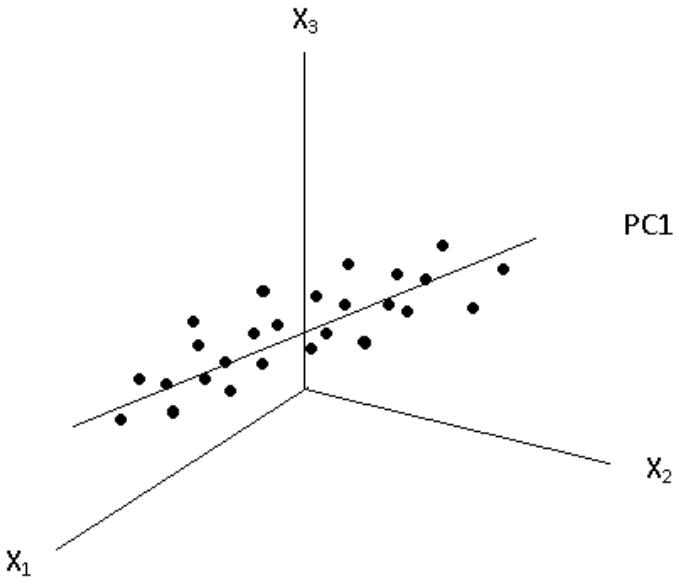


Figure 1.5. Hypothetical example of a principal component (reproduced from [73]).

The maximum number of PCs that can be derived from the independent variables is either the number of variables or one less than the number of samples, whichever is less.[73] For example, if spectral data, containing 1000 variables are obtained on 50 samples, the maximum number of PCs that could be used to create the new coordinate

system would be 49. In most cases, however, the actual number of PCs used to explain the variance is much smaller than the maximum number. PCA is typically used to analyze data and find patterns, and determine which variables contribute to the observed pattern. PCA forms the basis for several commonly used algorithms, such as soft independent modeling of class analogy (SIMCA), principal component regression (PCR), and partial least squares (PLS).

SIMCA is used for pattern recognition and classification. It is a soft modeling technique, in that it is possible for a sample to be assigned to more than one class.[72,78] SIMCA involves two stages, training and classification. During the training stage, PCA is carried out on the data set to make individual principal component models for each of the classes present within the data (Figure1.6).

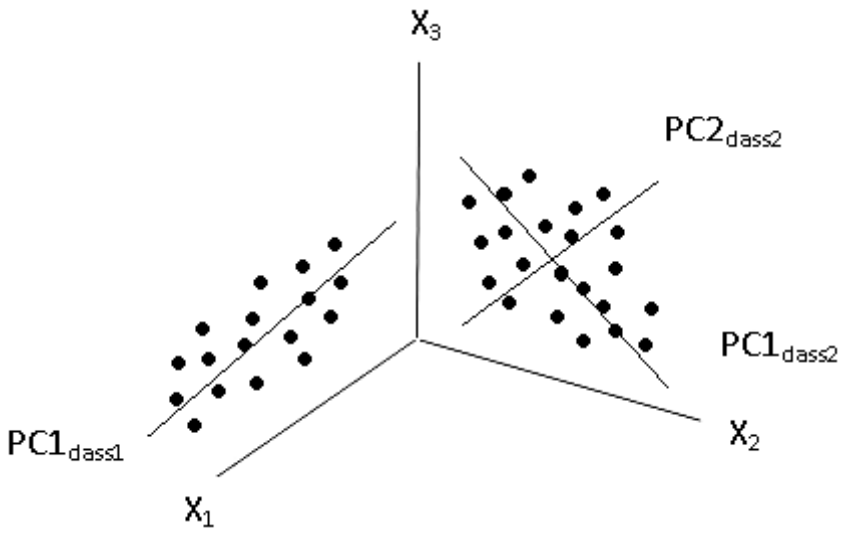


Figure 1.6. SIMCA: Construction of separate data classes.

The number of PCs used for each model may be different and depend on the variance within each class. The mathematical relationship between the independent

variables and the PCs is subsequently used to create a PC "map".[79] During the classification stage, new samples are projected onto this map to determine which established class they belong to. A sample is assigned to a given class if it occupies the space allocated to that class. If a sample falls outside the class borders of the established classes, it will not be assigned to any class. SIMCA is often used to identify and characterize subgroups within a group of samples (e.g. characterization of poultry meat in different stages of freshness), assign a new sample to an established class, or identify samples that are dissimilar compared to a known standard.

Several studies have demonstrated the use of SIMCA and PCA for pattern recognition or classification of data.[80-84] For example, Lonni and coworkers demonstrated that SIMCA can be successfully used to build a chemometric model for classification of three species of *Baccharis* plant, using their HPLC-DAD chromatographic data.[84] Principal component analysis was applied to chromatographic data of 74 alcoholic extracts of three *Baccharis* species to construct three class models, one for each species. Subsequent assignment of eight independent samples to each of the three classes was successfully achieved using the developed models.

As mentioned previously, PCA explains variance that is associated with the independent data set. If the goal is to construct a chemometric model for prediction of the dependent variable (Y), then MLR can be carried out on the new PCs to construct a prediction model. This two-step process is referred to as principal component regression (PCR).[72] An advantage of PCR over MLR is that it eliminates the issue of collinearity by using orthogonal PCs rather than directly correlating each X variable with the Y variable. Unfortunately, PCR has some drawbacks associated with it. As mentioned above, PCR is

a two step process, where PCA is carried out on the independent variables, followed by MLR analysis. The new coordinate system is created using PCs drawn in the direction of maximum variance through the *independent* variables. This means that each PC is created based on the amount of variance it explains in the X data, without any regard as to how that variance is correlated with the Y variable. For this reason, PCA is called an unsupervised learning technique.[73] The first few PCs include the most amount of signal regarding the X data with each subsequent PC including less signal and more noise. Therefore, when PCR is carried out, only the first few PCs are used, to incorporate mostly signal and little noise into the model. This can be an issue, given that some of the excluded PCs may contain underlying relevant information regarding the Y variable. Accordingly, to obtain the best prediction model, PCR must be carried out on different combinations of all PCs created to determine the optimum combination to use. This can be a tedious and time-consuming process, given that the total number of PCs created can be rather large. This is the main disadvantage of PCR.

Partial least squares (PLS) regression is another tool for reducing collinearity in independent variables and constructing models that can predict the value of the Y variable. In PLS, the new PCs are extracted from the independent variables based on their correlation with the Y variable.[72] In this way, each PC explains variance in both the X and Y data. As with PCA, the new PCs are created in the order of relevance. Since each PC explains information about both X and Y variables, all that is necessary to obtain the best prediction model is to use the right number of consecutive PCs to construct the model.[74] This eliminates the need for repeated calculations using different combinations

of PCs, as necessary with PCR. The chemometric model is constructed in the form of a regression vector:

$$\hat{y} = b_o + b_1x_1 + b_2x_2 + \ldots + b_nx_n \quad (1.1)$$

where y is the dependent variable, x_n is the independent variable, and b is the regression coefficient. There are different types of PLS algorithm available, depending on the nature of the X and Y variables used for modeling. PLS-1 is used to correlate two-dimensional independent data (e.g. absorbance vs. wavelength) with a single Y variable (e.g. concentration), while PLS-2 is used to correlate two-dimensional independent data with two or more Y variables (e.g. concentration of multiple compounds in a given sample). Tri-PLS is an N-way chemometric method which utilizes three-dimensional independent data (e.g. 3D chromatographic data) to build a regression model. As with PLS, there are two types of Tri-PLS algorithm. Tri-PLS1 can correlate the independent variables with a single Y variable, while Tri-PLS2 can correlate the independent data with two or more Y variables.

Once the chemometric model is constructed, it can be validated by predicting the value of the Y variable for a set of new samples, using their independent data alone. The validation step can be used to evaluate the model's predictive accuracy. Once validated, root mean square error of prediction (RMSEP) can be calculated for the validation samples, using the equation:[72]

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(y-yi)^2}{n}} \quad (1.2)$$

where y is the predicted concentration, y_i is the measured concentration for the *i*th validation sample, and n is the total number of validaion samples.

Several studies have successfully demonstrated the use of PLS analysis for construction of predictive models.[76,85-87] In a study by Galtier and coworkers, it was demonstrated that PLS can be used to build a chemometric model, correlating the near-infrared (NIR) spectra of 102 extra virgin olive oil samples to their fatty acid and triglyceride content, determined by gas chromatography (GC) and high performance liquid chromatography (HPLC).[85] The model was successful at predicting the fatty acid and triglyceride content for 23 validation samples, which were not part of the model calibration, based on their NIR spectra alone. In another study by Donmez and coworkers, the use of PLS modeling of HPLC-DAD chromatographic data was investigated for the simultaneous detection of four pharmaceutical compounds in cough syrup.[86] Twenty five synthetic syrup samples were prepared using various concentrations of potassium guaiacolsulfonate (PG), guaifenesin (GU), diphenhydramine HCl (DP), and carbetapentane citrate (CP). A chemometric model was constructed, correlating HPLC-DAD chromatographic data of the synthetic syrup samples with known concentrations of PG, GU, DP, and CP, using PLS-2. The developed model was successfully used to predict concentrations of the four compounds in a commercial syrup sample. Davis and coworkers demonstrated the use of PLS to explore correlations between ultraviolet (UV) spectra of 21 alcoholic extracts of habanero peppers and total concentration of two capsaicinoids, capsaicin and dihydrocapsaicin, present in the extracts.[76] The regression model was successful at predicting total concentration of capsaicin and dihydrocapsaicin in 10 validation samples based on their UV spectra, and 12 more validation samples five months after model calibration. In another study, Sikorska and coworkers successfully presented a method for simultaneous analysis of riboflavin (vitamin B_2), tryptophan,

tyrosine, and phenylalanine in beer, using chemometric modeling of fluorescence spectra.[87] As it can be seen, chemometrics is an effective tool for extracting useful information from the data and building predictive models.

It is important to note that real samples should always be used for constructing a chemometric model.[73] Artificial or laboratory-prepared samples cannot be used because they lack all possible compounds, unknowns, and interferences that are present in real samples that may influence the overall matrix. In complex mixtures, it is not possible to determine or quantitate all unknowns present in the mixture. Therefore it is not possible to design a sample that would be a true representative of future samples in the field. Consequently, if a model is constructed with artificial samples, it can only be applied to samples with known, controlled compositions, and will be unable of accurately predicting the dependent variable for real samples that may contain unknowns or interferences.[88]

Current Roadblocks to Cost-effective Production of Bioethanol

The technology for conversion of lignocellulosic biomass to ethanol has advanced considerably in recent decades. However, the lack of a firm understanding regarding process fundamentals, specifically, identification of inhibitory compounds produced during pretreatment, presents a barrier to cost-effective production of ethanol. Numerous studies have investigated the toxicity of several degradation products produced during pretreatment of biomass. However, most have employed the bottom-up experimental design, during which, one or more compounds are added to fermentation media and their effects on cell growth and ethanol production are observed. While this approach may be useful for assessing the relative toxicity of individual compounds, it does not take into account several factors contributing to overall hydrolysate toxicity. For instance, the

inhibitory effects of degradation products may not only be due to their individual concentrations, but also through synergistic inhibition. Furthermore, hydrolysates produced from different feedstock types, pretreatment methods and severities contain various types and concentrations of degradation products,[13,16] which can affect overall toxicity. Additionally, most toxicity studies have analyzed compounds that have already been identified. While many compounds have been identified to-date, hundreds of un-identified components are present in biomass hydrolysate. Any of these compounds may be inhibitory to downstream fermentation processes. Also, protocols used in various studies are inconsistent, measuring different output variables and using different fermentation procedures, inhibitor concentrations and/or fermentative organisms.[51,89-92] Therefore, comparing results between studies is often difficult. For example, in a study by Tran and Chambers,[89] it was reported that the relative toxicity of furfural was low, while in a study by Pfeifer and coworkers,[91] it was reported that the toxicity of furfural was significant. In addition to the disadvantages presented by the bottom-up approach, protocols used for toxicity tests are time consuming and labor intensive, making them unattractive to researchers exploring a broad spectrum of bioprocessing conditions. Therefore, traditional methods aimed at assessing toxicity or synergistic effects of certain compounds are unlikely to result in identification of all inhibitory constituents on a timescale that is consistent with current mandates for commercial production of cellulosic ethanol (Energy Independence and Security Act of 2007).[93] Consequently, there remains a need for development of a rapid and accurate method for assessment of hydrolysate toxicity and identification of inhibitory hydrolysate constituent.

Scope of the Dissertation

The primary goal of this dissertation was to develop two chemometric models for rapid prediction of hydrolysate fermentability and identification of fermentation inhibitors, respectively. Compositional data obtained on biomass hydrolysates were correlated with fermentability to build the models. Fermentability was defined as percent inhibition of ethanol production compared to a control. The proposed methodology is a more rapid, top-down approach, in that it utilizes compositional information regarding whole hydrolysate, taking into account additive and/or synergistic effects from all constituents which may affect overall toxicity. The developed models provide significant improvements in throughput and labor when compared to traditional methods used for assessing hydrolysate fermentability and identification of inhibitory compounds. This work is expected to promote a better understanding of inhibitory constituents in biomass hydrolysate, which may facilitate the development of appropriate detoxification methods and/or inhibitor-resistant microorganisms that are capable of carrying out the fermentation process in the presence of inhibitory compounds. The developed approach can be applied to various feedstocks, pretreatment methods, and fermentative organisms. Given that a majority of studies related to biofuel production have been carried out on ethanologenic organisms, *Saccharomyces cerevisiae* yeast was selected for proof-of-concept purposes. Furthermore, corn stover and dilute-acid pretreatment were the selected feedstock and pretreatment method, respectively, given that they are commonly employed in studies related to production of bioethanol.

The first chemometric model was developed for prediction of hydrolysate fermentability using UV-visible spectroscopic data. The model was constructed,

correlating UV-visible spectral data of raw hydrolysate samples with percent inhibition values. Percent inhibition was determined using traditional batch fermentation methods. Details regarding the development of this model are discussed in Chapter 2 of this dissertation. The developed model demonstrated high prediction accuracy for a set of validation samples that were compositionally different from samples used for model construction. Additionally, four spectral regions were identified by the model as having the strongest correlation with microbial inhibition. The developed chemometric approach offers significant improvements in throughput and labor when compared to traditional batch fermentation methods.

A second chemometric model was developed, correlating HPLC-UV chromatographic data of raw hydrolysates with their percent inhibition. Detection was monitored at the four wavelengths identified by the UV-visible model as the most significant spectral regions. Details regarding the development of this model are discussed in Chapter 3 of this dissertation. The goal was to demonstrate that chemometric modeling can be used to correlate microbial inhibition with HPLC-UV chromatographic data, and to identify significant retention times with the highest correlation with inhibition. Two retention times were identified by the model as having the strongest correlation with inhibition. The identified retention times corresponded to compounds that are known inhibitors of fermentative microorganisms. To determine whether better resolution or more universal detectability of sample constituents may lead to identification of more compounds, all hydrolysate samples used for model construction were analyzed by ion chromatography (IC) with conductivity detection. Another chemometric model was developed, correlating IC data with percent inhibition. Details

of IC analysis and model development are also discussed in Chapter 3. Conclusions and final remarks of this dissertation are discussed in Chapter 4, as well as future perspective on the research presented.

CHAPTER TWO

Prediction of Hydrolysate Fermentability Using Chemometric Modeling of UV-Visible Spectroscopic Data

Introduction

Production of fuel ethanol from lignocellulosic biomass holds a broad range of economical and environmental benefits and has thus been the subject of extensive research.[8-10] Lignocellulosic biomass is composed of cellulose, hemicellulose, and lignin, with the former two composing up to 70% of the fermentable sugars.[63,94,95] Generally, a pretreatment step is required to breakdown lignin and hemicellulose and render cellulose accessible for further enzymatic hydrolysis.[10,13,27,29] However, under pretreatment conditions, a variety of degradation products are formed that are potentially inhibitory to downstream fermentation processes.[14,46,96-98] Batch fermentation methods are typically used to assess the toxicity of pretreatment hydrolysate.[40] These methods are extremely time-consuming and labor intensive, which makes them unattractive to researchers exploring a broad spectrum of bioprocessing parameters (i.e. feedstock, pretreatment chemistry and severity, etc.). Consequently, there remains a need for development of a rapid and accurate method for assessment of hydrolysate toxicity.

Some progress has been made towards an alternative approach for assessment of hydrolysate toxicity. Larsson and coworkers conducted an extensive study, investigating the influence of dilute-acid pretreatment severity (i.e. time, temperature, and acid concentration) on the release of sugars and by-products and on ethanol formation using the combined severity factor (CSF).[15] The CSF is represented by:[99]

$$\text{CSF} = \text{Log } R_o - \text{pH} \quad (2.1)$$

$$R_o = t \times \exp\left[(T_r - T_b)/14.75\right]$$

where t is residence time, T_r is the reaction temperature and T_b is a reference temperature (100°C). The pH is calculated from the amount of sulfuric acid used for pretreatment. Seventy six different severity conditions were chosen, corresponding to a CSF range of 1.4-5.4. Fermentability of hydrolysates, as measured by ethanol yield and productivity, generally decreased with increasing CSF, with the optimum CSF values corresponding to a range of 2.9-3.1. However, not all samples followed this trend. In fact, the hydrolysate sample with the highest ethanol yield had a CSF value of 3.4. In another study carried out by Chen and coworkers, the relationship between reaction severity and accumulation of potentially inhibitory degradation products was assessed.[100] Twelve corn stover hydrolysates were generated, covering a time and temperature range of 2-64 minutes and 160-200°C, respectively. The relationship between the concentration of nineteen degradation compounds and reaction severity was assessed. The results demonstrated that at constant CSF, total concentration of degradation compounds varied significantly depending on the time and temperature employed. Furthermore, in some cases, a higher concentration of degradation compounds was observed at lower CSF values. These results further confirmed that the severity function cannot serve as an appropriate predictor for accumulation trends for degradation compounds.

In a recent work by Morita and coworkers,[101] the use of near infrared (NIR) spectroscopy was investigated for screening recombinant *Saccharomyces cerevisiae* strains and evaluating their fermentation performance in YPD media (yeast extract-peptone-dextrose). Four different recombinant xylose-fermenting strains (one diploid

and three haploid) were used in the study. Supernatant samples were collected from fermentation solutions and analyzed using NIR spectroscopy. Samples were also analyzed for the amount of consumed glucose and xylose, produced ethanol, glycerol, and xylitol to evaluate fermentation performance. A classification model was constructed using soft independent modeling of class analogy (SIMCA) to investigate the difference in spectral data among the strains. Partial least squares was used to construct a separate prediction model for each strain, correlating the NIR spectral data with the concentration of consumed glucose and xylose, produced ethanol, glycerol, and xylitol. In addition, one prediction model was constructed, using data from all four strains. Overall, the results demonstrated that chemometric modeling of NIR spectral data can serve as a rapid tool for evaluating fermentation performance of recombinant strains. It is important to note that since fermentations were carried out in YPD media, the performance of recombinant strains may vary when tested in real hydrolysate samples. Chemometric modeling of NIR spectral data for prediction of hydrolysate fermentability was previously investigated in our laboratory. However, preliminary experiments demonstrated a poor correlation between NIR spectral data and fermentability for the hydrolysate samples studied. Therefore, focus was turned to other spectroscopic methods, such as UV-visible spectroscopy.

The use of UV absorption spectroscopy for monitoring furans in dilute acid hydrolysate has been investigated previously by Martinez and coworkers.[42] Their study also investigated the use of furan content as a predictor of hydrolysate toxicity. Fourteen hydrolysates were fermented, and the amount of ethanol produced was plotted as a function of furan content. For 11 hydrolysates, maximum ethanol production correlated

with a furan content below 0.2 g/L. For three hydrolysates, however, maximum ethanol production was obtained at a furan content below 0.5 g/L. The difference observed was considered to be due to a change in the ratio of 5-hydroxymethylfurfural (HMF) to furfural. Hydrolysates of lower toxicity had an HMF to furfural ratio of 0.8-1.1, while hydrolysates of higher toxicity had an HMF to furfural ration of 0.15-0.2. It was concluded that total furan content can serve as a predictor of hydrolysate toxicity, given that "there are not large differences in the proportions of HMF and furfural".[42] To investigate the correlation between UV spectral data and furan content, UV analyses were carried out on 14 bagasse hydrolysate samples and the difference in spectral response at 284 and 320 nm was plotted as a function of total furan content (furfural and HMF), determined by HPLC-UV. Although good correlation was observed between single wavelength absorption data and furan content, the univariate nature of this approach does not account for variable changes in the amounts of other absorbing compounds, which could impact the accuracy of the method. Furthermore, the authors stated that hydrolysate toxicity is caused by other compounds in addition to furans. Consequently, a more comprehensive, full-spectrum technique is needed for rapid assessment of hydrolysate inhibition.

Herein we present a proof-of-concept chemometric model for prediction of hydrolysate fermentability using UV-visible spectral data. Twenty one corn-stover hydrolysate samples of varying pretreatment severities were used to develop the model. UV-visible spectroscopic analyses were carried out on raw hydrolysates in the spectral range of 190 to 450 nm. Fermentability (i.e. percent inhibition of ethanol production) was determined for each hydrolysate using traditional batch fermentation methods.

Partial least squares (PLS-1) was used to regress spectral data against inhibition values and construct the chemometric model. Following model construction, validation was carried out by predicting percent inhibition of five hydrolysate samples, different from ones used to construct the model. Predicted inhibition values were compared with experimentally determined inhibition values to assess the model's predictive accuracy. Following validation, the model was used to identify spectral regions with the highest correlation with microbial inhibition. The chemometric approach utilizes compositional information regarding whole hydrolysate for assessment of toxicity, taking into account additive and/or synergistic effects from all constituents. The novelty of this work is in demonstrating that rapid and accurate prediction of microbial inhibition is feasible using chemometric modeling of UV-visible spectroscopic data. The developed approach offers significant improvements in throughput and labor when compared to traditional batch fermentation methods.

Experimental: Materials and Methods

Reagents and Standards

Analytical grade glucose, yeast extract, peptone, ethanol, sulfuric acid and citric acid monohydrate were purchased from Sigma-Aldrich (St. Louis, MO, USA). Five ethanol standards at concentrations of 10, 4, 2, 1, and 0.2 mL/L were prepared from an ethanol (100 mL/L) stock solution. These standards were subsequently used to build an external standard calibration curve for HPLC analyses of ethanol concentration. Distilled water was purified and deionized to 18 MΩ with a Barnstead Nanopure Diamond UV water purification system.

Microorganisms

Saccharomyces cerevisiae D_5A was obtained from National Renewable Energy Laboratory (NREL), (Golden, CO, USA). The frozen stock culture was prepared using standard NREL protocols for *S. cerevisiae*.[102]

Preparation of Hydrolysates

Corn stover (4.75% moisture content [(w/w)]) was provided by NREL in 1-2 mm mesh size. Dr. Peter van Walsum (University of Maine, Forest Bioproducts Research Initiative) provided the hydrolysate samples, which were generated using a Dionex accelerated solvent extractor (ASE 300). Twenty six pretreatments of varying severities were carried out in 0.7% (w/w) sulfuric acid at a solids concentration of 8% (w/v). Table 2.1 lists the pretreatment severities used to generate each hydrolysate sample. It was anticipated that this range of severities would be representative of future samples to be analyzed using the developed model. In addition, it was anticipated that the selected severity range would provide variable amounts of degradation products, which can have varying effects on inhibition.[13] Once received, the pH of the hydrolysate was adjusted to 4.85 using calcium hydroxide, and particulates were removed by filtration with 0.45-µm membrane filters prior to fermentation.

Fermentation of Hydrolysates

Hydrolysate samples were batch fermented in triplicate using the fermentation protocol specified in NREL Laboratory Analytical Procedure (LAP) Lignocellulosic Biomass Hydrolysis and Fermentation.[102] No enzymatic hydrolysis was carried out prior

Table 2.1. Pretreatment parameters used to obtain calibration and validation samples.

Sample number[a]	Time (min)	Temperature (°C)	CSF[b]
1	16	190	2.70
2	64	170	2.72
3	24	185	2.73
4	50	175	2.76
*5	14	195	2.79
6	40	180	2.81
*7	60	175	2.84
8	22	190	2.84
*9	64	175	2.86
10	12	200	2.87
*11	35	185	2.90
12	26	190	2.91
13	55	180	2.95
14	20	195	2.95
15	60	180	2.98
16	22	195	2.99
17	16	200	3.00
18	24	195	3.03
*19	18	200	3.05
20	55	185	3.09
21	40	190	3.10
22	30	195	3.12
23	24	200	3.17
24	50	190	3.20
25	45	195	3.30
26	64	190	3.31

[a] Validation samples are represented by asterisk

[b] Corn stover hydrolysates were generated using ASE 300, 0.7% (w/w) H_2SO_4, 8% (w/v) solids concentration

CSF = log [t x exp $(T_r - T_b)/14.75$] − pH

to fermentation. The fermentation inoculum was prepared by transferring one thawed stock vial of *S. cerevisiae* to a sterile YPD (yeast extract-peptone-dextrose) flask, and incubated for 10-14 hours at 30°C with rotational agitation at 135 rpm. Fermentations were performed in 150-mL flasks, containing 50 mL of total working solution. The working solution included 11 mL of hydrolysate, 5 mL of sterilized 10x concentrated yeast-peptone (10x-YP) stock solution, 5 mL of sterilized glucose stock solution (500 g/L) and 29 mL of sterilized DI water. The 10x YP stock solution consisted of 100 g/L yeast extract and 200 g/L peptone in water and its pH was adjusted to 4.85 using sterilized citrate buffer. The glucose stock solution was used to adjust the final concentration of fermentable sugars to 50 g/L. The concentration of fermentable sugars in hydrolysate samples was determined by high-performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) using a previously established methodology.[103] Control samples only contained 5 mL of sterilized growth medium, 5 mL of sterilized glucose stock solution, and 40 mL of sterilized DI water. All constituents were added aseptically to sterilized fermentation flasks, which were subsequently capped with rubber stoppers and an aluminum seal. Prior to inoculation, the yeast inoculum was checked for contamination and analyzed according to NREL procedure.[102] The optical density (O.D.) of the inoculum was measured at 600 nm. This value was used to calculate the amount of culture needed to inoculate each fermentation flask to a starting O.D. of 0.5. Following inoculation, samples were incubated at 30°C in a rotary shaker at 135 rpm for 48 hours. Samples of 0.5 mL were drawn prior to inoculation and at 0, 2, 4, 6, 8, 10, 12, 14, 17, 19, 24 and 48 hours after inoculation.

Percent inhibition was calculated by comparing the amount of ethanol produced for each sample at a given sampling time to the amount produced in the control.

To determine precision of experimental fermentations, a pooled standard deviation (s_{pool}) was calculated. This is a statistical tool for estimating variance when several series of measurements are taken under the same conditions, and it is assumed that the precision remains the same even if the observed values differ. Pooled standard deviation is calculated by:[104]

$$S_{pool} = \sqrt{\frac{\sum^{N_1}(X_i - \overline{X}_1)^2 + \sum^{N_2}(X_j - \overline{X}_2)^2 + \cdots + \sum^{N_t}(X_p - \overline{X}_y)^2}{N_1 + N_2 + N_3 + \cdots - N_y}} \quad (2.2)$$

2

where for a given sample, X_i is the measured value from the *i*th measurement, $\overline{X}_1$ is the mean of all measurements, N_1 is the total number of measurements made, and N_y is the total number of samples being combined. Additionally, the confidence interval for each sample was calculated at the 95% confidence limit, using:

$$CI_{95} = \frac{S_{pool} * z}{\sqrt{n}} \quad (2.3)$$

where *z* is the standard score and *n* is the number of measurements for a given sample.

High-Performance Liquid Chromatography Analysis

An HPLC external standard calibration curve was constructed using ethanol standard solutions. The calibration curve was subsequently used to determine ethanol concentration (mL/L) using a Dionex DX-600 series liquid chromatograph (Dionex, Sunnyvale, CA, USA). The HPLC system consisted of an AS50 auto-injector, GS50 Gradient pump, LC30 Chromatography oven and a Shodex RI-101 differential refractive index detector. Chromatographic separation was achieved using a 150 mm x 7.8 mm Aminex HPX-87H fast fermentation column with 9 µM particle size (Bio-Rad, Hercules,

CA, USA), operated at 60°C. A 30 x 4.6 mm IG Cation H guard column (Bio-Rad) was used to protect the analytical column. An isocratic eluent consisting of 32 mM sulfuric acid at a flow rate of 1.0 mL/min was employed to achieve chromatographic separation.

UV-visible Spectroscopic Analysis

Prior to fermentation, absorption spectra were collected using a PerkinElmer Lambda 35 Dual-beam Spectrometer (Perkin-Elmer, Waltham, MA, USA). Each hydrolysate sample was diluted 2000-fold in DI water and its spectrum was obtained over the range of 190 to 450 nm using a 1.0 cm path-length quartz cell.

Chemometric Analysis

Unscrambler 9.8 (CAMO Software Inc., Woodbridge, NJ, USA) was used for chemometric analyses. UV-visible spectra and mean (n=3) inhibition data were imported into the software and mean-centered prior to all chemometric analyses. Partial least squares (PLS-1) was used to regress inhibition values against spectral data. The method used to calibrate the model was leave-one-out cross-validation.

Five hydrolysate samples were selected for model validation and the remaining 21 samples were used to construct the model. Validation samples (5, 7, 9, 11, and 19) were selected by placing mean inhibition values of all 26 samples into five categories, with each category corresponding to a specific inhibition range (20-30%, 30-40%, 40-50%, 50-60%, and 60-70%). Subsequently, one sample was randomly selected from each of the five categories, using a random number generator. This process ensured that inhibition values of the validation samples spanned the range of inhibition values used for model calibration.

Results and Discussion

Inhibition of Ethanol Production

It is important to point out that the concentration of sugars released during pretreatment in the hydrolysate samples was negligible and that the amount of ethanol produced during fermentation was primarily from the glucose added (50 g/L) prior to fermentation. Since all hydrolysates had the same final concentration of glucose, the observed fermentation data do not enable an evaluation of actual ethanol yield from hydrolysates, but rather a relative ranking of hydrolysate toxicity. Figure 2.1A shows the mean concentration of ethanol produced in each sample at specific sampling intervals (0, 2, 4, 6, 8, 10, 12, 14, 17, 19, 24 and 48 hours) following inoculation. Even though all samples show a similar amount of ethanol produced after 48 hours, the rate of ethanol production was different for each sample. In order for calibration data to be representative of future samples, a sampling time that provided the greatest variation in inhibition was needed for modeling purposes. From careful inspection of data represented in Figure 2.1A, it was determined that the greatest variability in ethanol concentration was observed at the 14 hour sampling time. Accordingly, data from this sampling interval were selected for model development. Mean percent inhibition values at 14 hours are presented in Figure 2.1B. No direct correlation was observed between percent inhibition and CSF values listed in Table 2.1. For instance, samples 7 and 8 have identical CSF values of 2.84, but mean percent inhibition values of 71 and 46, respectively.

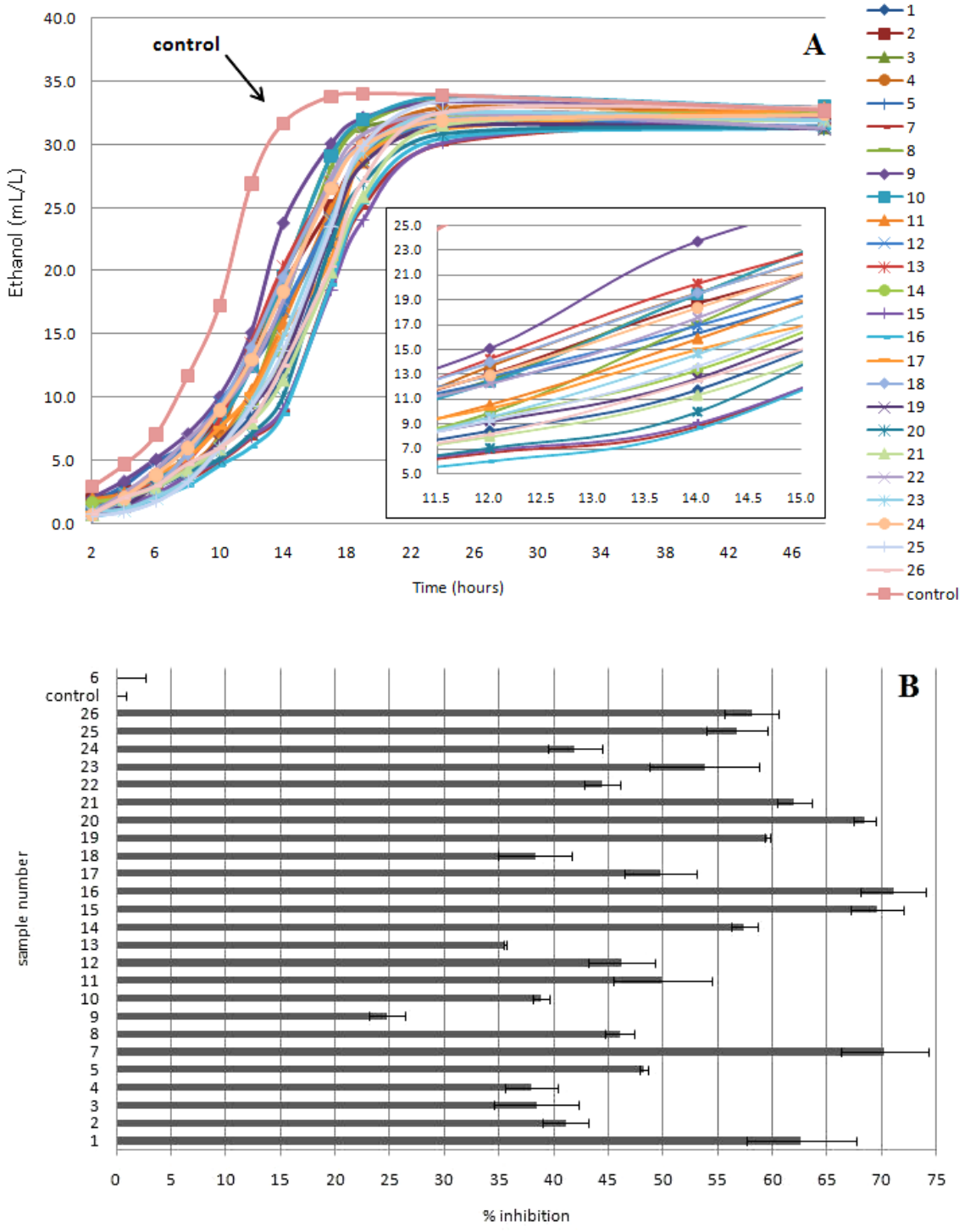


Figure 2.1. Observed ethanol concentration and percent inhibition for hydrolysate samples: A) Mean concentration values of ethanol produced from triplicate experiments as a function of sampling time (0, 2, 4, 6, 8, 10, 12, 14, 17, 19, 24 and 48 hours), B) Mean percent inhibition of ethanol production at the 14 hour sampling time. Sample 6 showed 0% inhibition at 14 hours. Error bars represent one standard deviation.

UV-Visible Spectroscopic Data

UV-visible spectra of the 26 hydrolysate samples are depicted in Figure 2.2A. Most notably in this data set, maxima are observed at 223 nm and 279 nm, with the latter maxima being the most prominent. UV maxima at 279 are characteristic of furans, specifically furfural and HMF, which have long been considered among the most inhibitory degradation products in biomass hydrolysates.[33,42]

Figure 2.2B is a plot of the mean-centered spectral data, which was constructed by averaging the spectra on a wavelength-by-wavelength basis then subtracting the averaged spectrum from each individual spectrum.[76] Mean-centering is most useful when the variance is of interest and not necessarily the signal of the data. Accordingly, a mean-centered plot can be used to identify regions with high variation in the data. Figure 2.2B shows that most of the variance is occurring below 300 nm with several local maxima occurring between 200 and 280 nm. Given that the model focuses on the variation in spectral data that is correlated with the observed variation in inhibition data, it is likely that the change in absorbance at these wavelengths has a strong correlation with microbial inhibition.

Chemometric Model Construction and Validation

UV-visible spectral data of the 21 calibration samples and the inhibition data were imported into CAMO software and mean-centered prior to calibrating the regression model. Partial least squares (PLS-1) was used to develop the model. Partial least squares focuses on variance in the independent data set that is correlated with the dependent variable while minimizing the influence of uncorrelated variables.[73] The model was constructed in the form of a regression vector:

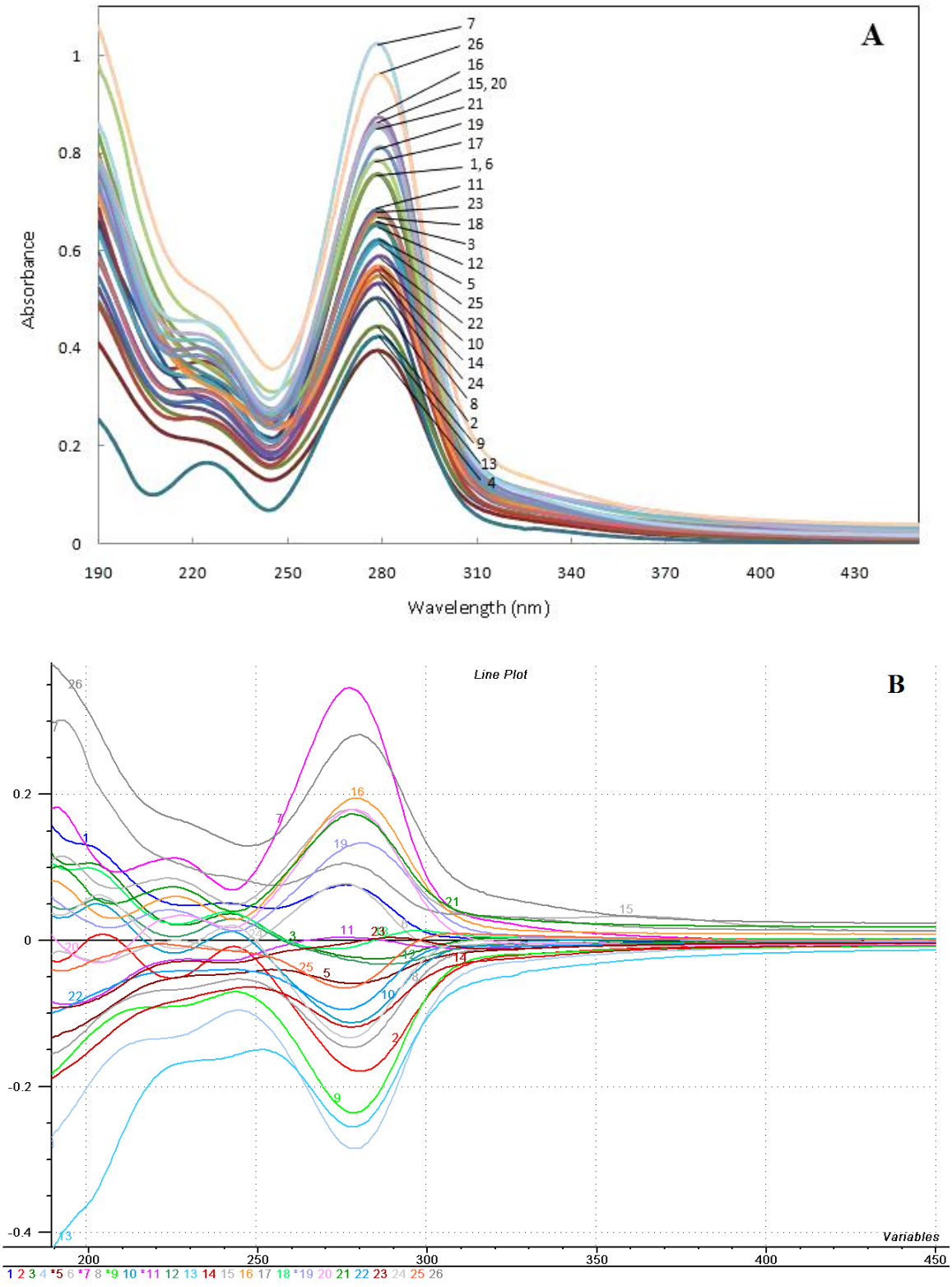


Figure 2.2. A) UV-visible spectra of raw hydrolysate samples, B) mean-centered spectra of hydrolysate samples.

$$\hat{y} = b_o + b_1x_1 + b_2x_2 + \ldots + b_nx_n \quad (2.4)$$

where $\hat{y}$ is inhibition for a given sample, b is the regression coefficient determined by the PLS-1 algorithm, and x_n is the measured absorbance at the selected wavelength (λ_1 to λ_n). The chemometric model was calibrated using leave-one-out cross-validation.[73,105] During this process, one calibration sample was kept out and a model was computed using the remaining samples. The constructed model was subsequently used to predict inhibition for the sample that was left out. This process was repeated until every sample was left out of calibration once. The difference between predicted values and experimentally determined values of all calibration samples were used to calculate the predicted residual error sum of squares (PRESS):

$$\text{PRESS} = \sum (y - y_i)^2 \quad (2.5)$$

where y is the predicted inhibition and y_i is the experimentally determined inhibition. This process was repeated for models using a different number of principal components (PCs). The calculated PRESS was subsequently averaged over all calibration samples to calculate the residual validation variance.[106] Figure 2.3A is a plot of residual variance versus the number of PCs used to explain the model.

The residual validation variance plot represents the error that is expected with future predictions, depending on the number of PCs used. To obtain the best prediction model, it is necessary to use the maximum number of components that provide the lowest value of Y-variance. Utilization of additional PCs introduces noise into the model, and leads to an increase in prediction error. Inspection of data in Figure 2.3A revealed that use of two PCs resulted in minimal residual variance. Figure 2.3B shows the scores plot, which displays the location of samples in the coordinate system for the first two PCs. It

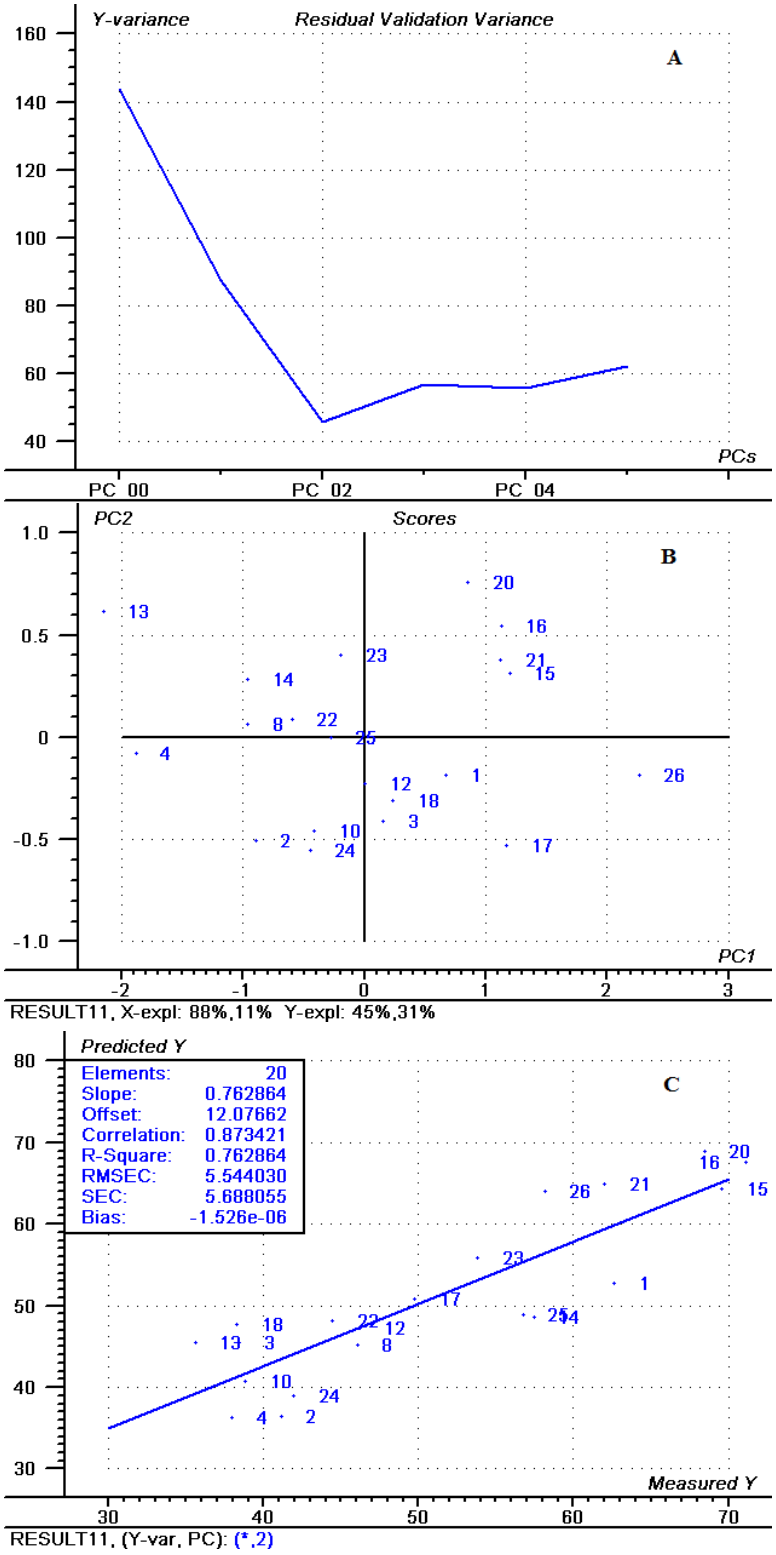


Figure 2.3. Regression overview of model constructed with full UV-visible spectral data (190-450 nm): A) residual variance plot, B) scores plot, C) predicted vs. measured plot.

is demonstrated below the x-axis that the first two PCs collectively explain 99% of the variance in the spectral data, and 76% of the variance in inhibition data, respectively. For the model developed here, most samples are well-distributed across the coordinate system. It should be noted that sample 6 was identified as an outlier using the software. This may be due to the fact that sample 6 presented 0% inhibition at the 14 hour sampling time. Omission of sample 6 from the calibration sample set significantly improved the model's correlation.

Following model calibration, the developed regression vector was used to predict inhibition for every calibration sample (n=20). The predicted values were then plotted against measured values to obtain the predicted vs. measured plot (Figure 2.3C). This plot is used to evaluate a model's fit. A perfect model would have an R^2 of 1. For the developed model, R^2 was 0.76. It should be noted that R^2 for the model constructed prior to omission of the outlier (sample 6) was 0.19. For the developed model, the offset (intercept) was 12. Root-mean-square error of calibration (RMSEC), standard error of calibration (SEC), and bias (systematic error) were 6, 6, and -2 E^{-6}, respectively, indicating negligible systematic error and reasonable accuracy. The correlation coefficient is the square-root of R^2 and its sign indicates a positive or negative correlation. Thus, the observed correlation coefficient of 0.87 demonstrates that an increase in inhibition is expected with an increase in absorption. While the correlation coefficient and R^2 provide a measure of the model's fit, the best way to evaluate its predictive ability is to validate it using new samples.

Validation of the model was carried out with five hydrolysate samples that were not used to construct the model. UV-visible spectra of validation samples were imported

into the software and used by the developed regression vector to predict inhibition values. Table 2.2 shows the results of the validation step, including experimental and predicted inhibition for each validation sample along with root-mean-square error of prediction (RMSEP).

Table 2.2. Prediction results for full-spectrum and four-wavelength models.

Validation sample number[a]	Experimental % inhibition[b]	Predicted full-spectrum model[c]	Predicted four-wavelength model[d]	Absolute error
5	48 ± 6	52	52	4
7	70 ± 6	75	75	5
9	25 ± 6	36	36	11
11	50 ± 6	55	55	5
19	60 ± 6	63	63	3
RMSEP		6	6	

[a] pretreatment parameters for validation samples are presented in Table 1. [b] mean percent inhibition from triplicate fermentations of hydrolysate and corresponding 95% confidence intervals. Pooled standard deviation was 5. [c] Predicted inhibition using model constructed with full UV-visible spectral data (190-450 nm). [d] Predicted inhibition using model constructed with spectral data at 201, 223, 242, and 279 nm.

Four of five samples (5, 7, 11, and 19) had predicted values that were within 95% confidence limits of experimentally determined inhibition. Sample 9 had a predicted inhibition value that was not within the confidence limits. This may be due to the fact that the experimentally determined inhibition value for sample 9 was outside of the range of values used for model calibration. Predicted values for all validation samples had a high bias associated with them. An explanation for the observed bias is unclear at this point. Overall, the results confirmed that accurate prediction of microbial inhibition is

feasible using chemometric modeling of UV-visible spectral data, provided that inhibition values for future samples fall within the range of values used for calibration.

Identification of Significant Spectral Regions

Once the model was successfully validated, two chemometric tools, regression coefficient plot and x-loadings, were used to identify spectral regions with the strongest correlation to microbial inhibition. Figure 2.4A is a plot of the different *b* coefficients from the regression vector (equation. 2.1) as a function of wavelength. Maxima in this figure (201, 223, 242, and 279 nm) represent spectral regions where the change in absorbance is strongly correlated with the observed variation in inhibition. Identification of these wavelengths as significant variables was also supported by the x-loadings plot (Figure 2.4B). The loading is the cosine of the angle between each variable and the corresponding PC.[74] The higher the loading value, the more significant that variable is to the PC. A loading value of 1 or -1 (angle of 0 or 180 degrees) indicates that the variable is parallel with the component and contributes a significant amount of information to it. On the other hand, a loading value of 0 (angle 90 degrees) indicates that the variable contributes little information to that component. In Figure 2.4B, peak maxima were identical to the ones identified in Figure 2.4A.

Investigation of the Significance of Identified Wavelengths

To investigate the relative significance of the four wavelengths identified (201, 223, 242, and 279 nm), a second multivariate model ("four-wavelength model") was constructed, correlating spectral data at the identified wavelengths with inhibition data, and the results were compared with the previously developed full-spectrum model. The

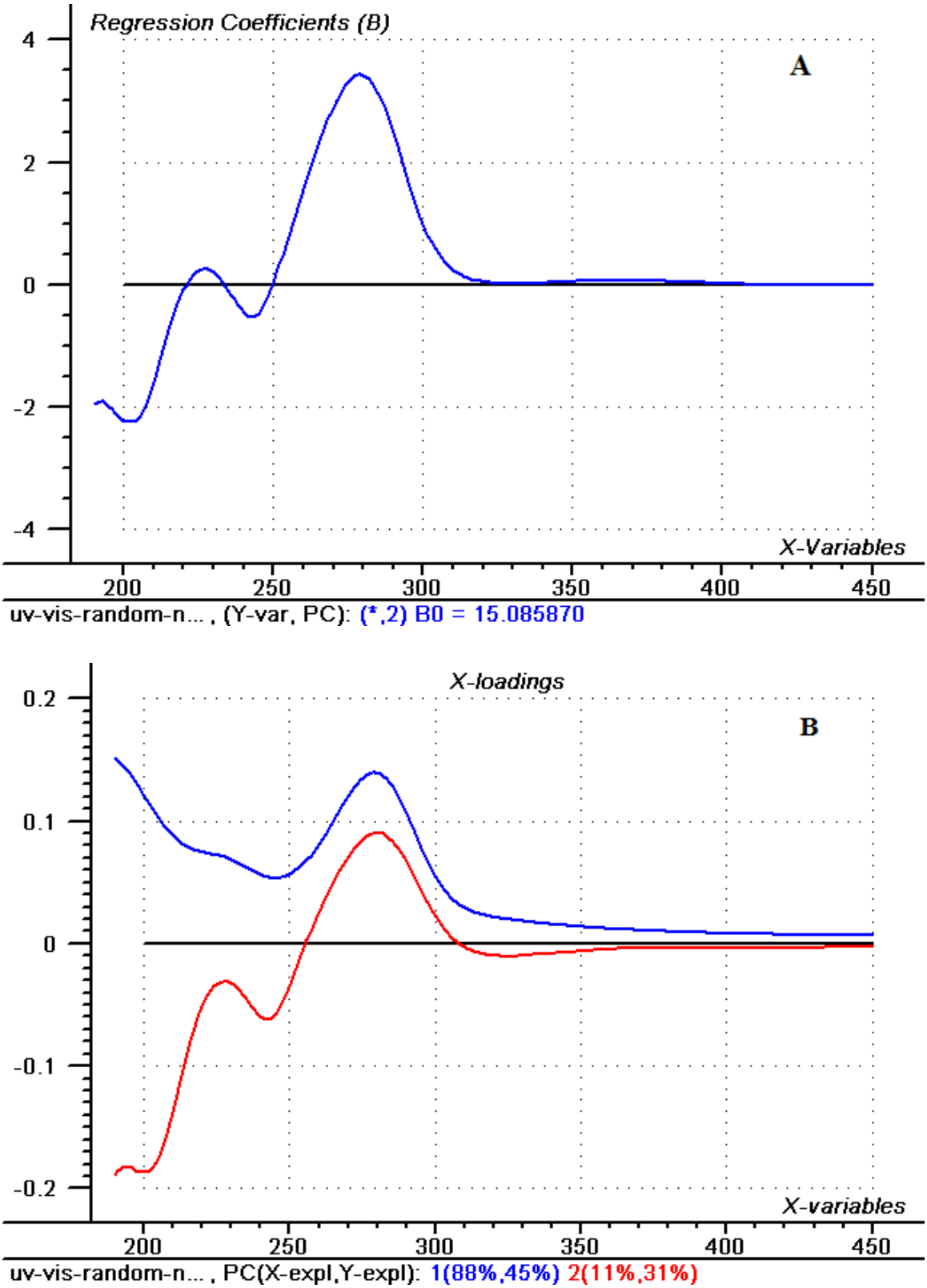


Figure 2.4. A) Regression coefficient plot for UV-visible model, B) X-loadings plot for the first two principal components.

same hydrolysate samples were used to construct and validate the four-wavelength model as the full-spectrum model. The regression overview for the four-wavelength model is presented in Figure 2.5. Validation of the model was carried out by predicting inhibition of five validation samples, using their spectral data at 201, 223, 242, and 279 nm. Results of validation are presented in Table 2.2. The four-wavelength model had prediction results that were identical to the full-spectrum model. For this model, R^2 and RMSEP were 0.76 and 6, respectively. Results indicate that the variation in absorbance at the four identified wavelengths is responsible for the observed variation in inhibition.

Assessment of Univariate Techniques for Prediction of Hydrolysate Fermentability

Several univariate models were constructed to evaluate the use of single wavelength absorption data for prediction of microbial inhibition. Univariate analyses were carried out by plotting the absorbance at each of the identified wavelengths (201, 223, 242, 279 nm) as a function of inhibition for the calibration samples used to construct the chemometric models. Data were subsequently fit to a straight line (Figure 2.6 and 2.7) and the corresponding linear equation for each univariate analysis was used to predict inhibition for the five validation samples. Results of the univariate analyses are presented in Table 2.3. For the single wavelength models, R^2 was lower than the multivariate models. Furthermore, RMSEP for the single wavelength models at 201, 223, 242, and 279 nm was 13, 10, 12, and 9, respectively. Thus, the 279 nm model the best univariate prediction model.

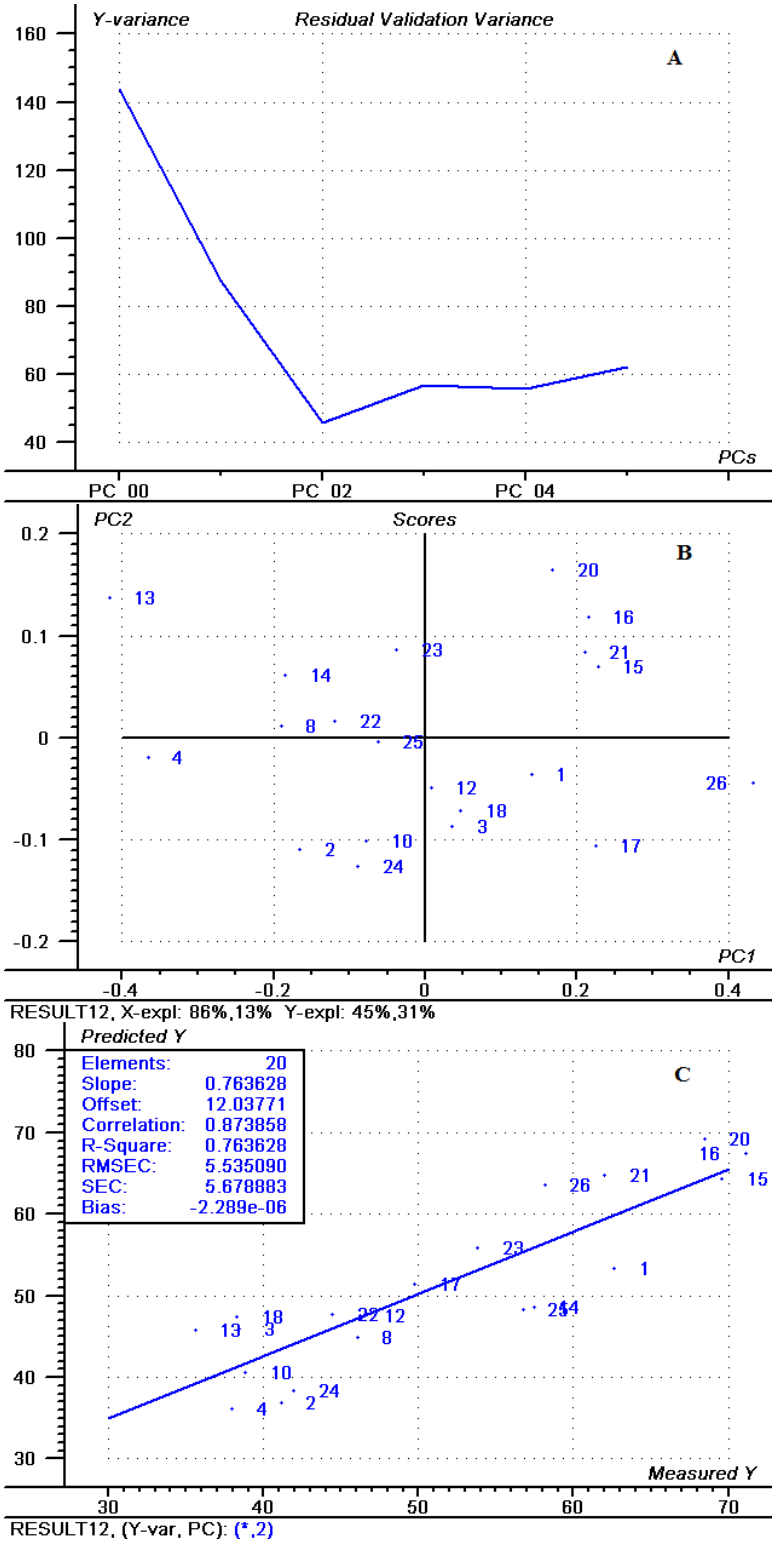
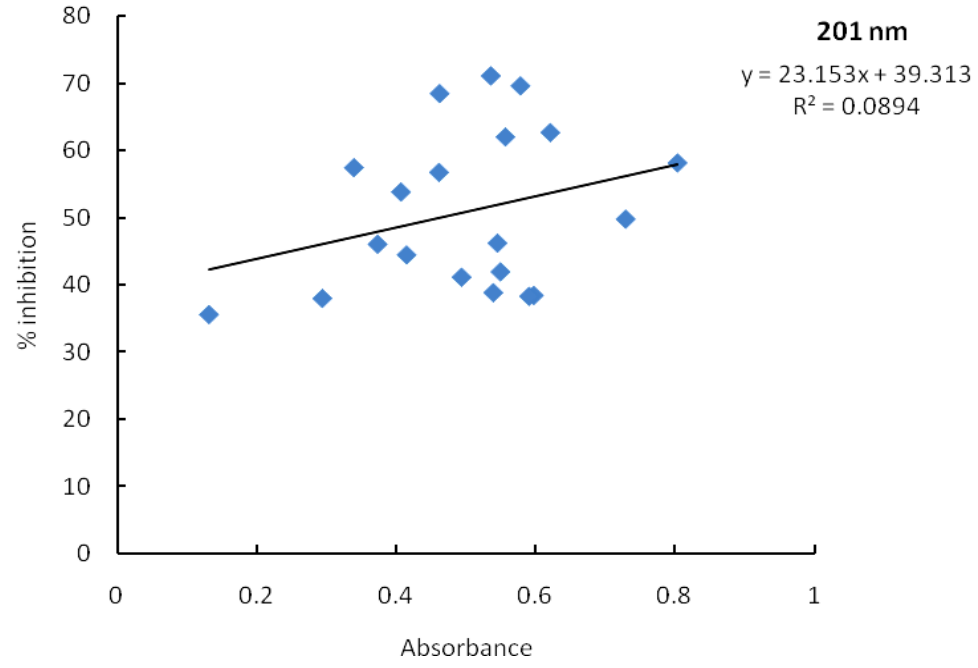


Figure 2.5. Regression overview for model constructed using spectra data at 201, 223, 242, and 279 nm: A) residual variance plot, B) scores plot, C) predicted vs. measured plot



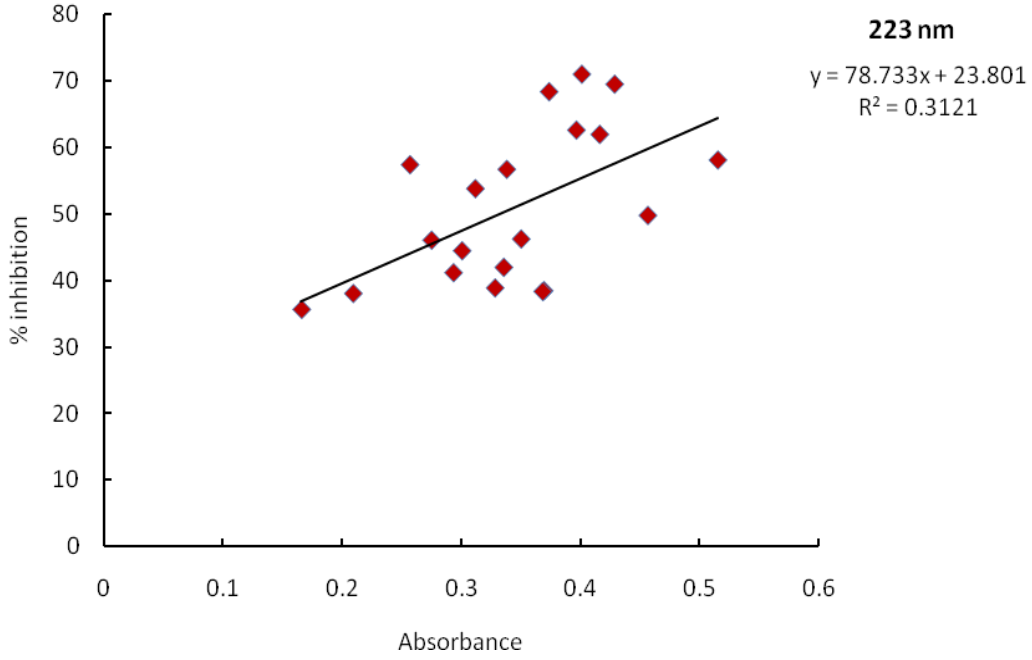


Figure 2.6. Plot of absorbance (at 201 and 223 nm) as a function of percent inhibition.

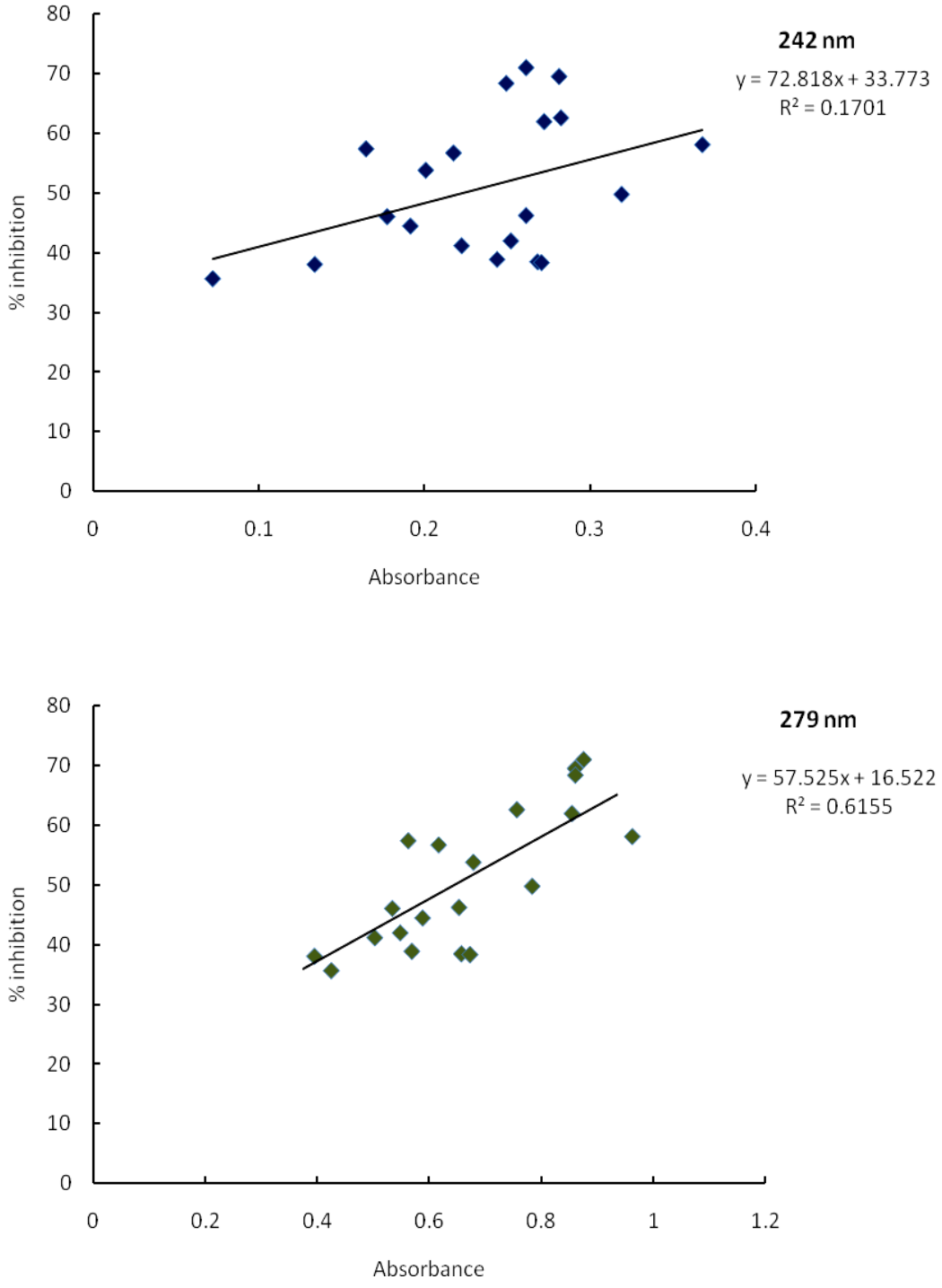


Figure 2.7. Plot of absorbance (at 242 and 279 nm) as a function of percent inhibition.

Table 2.3. Prediction results using univariate approach (single-wavelength models)

Validation sample number	Experimental % inhibition	Predicted[a] at 201 nm	Predicted[b] at 223 nm	Predicted[c] at 242 nm	Predicted[d] at 279 nm
5	48 ± 6	48	47	47	52
7	70 ± 6	54	60	56	75
9	25 ± 6	48	44	45	42
11	50 ± 6	49	49	49	56
19	60 ± 6	51	54	52	63
RMSEP		13	10	12	9

[a] single wavelength models at 201 nm ($y = 23.153x + 39.313$, $R^2 = 0.0894$), [b] 223 nm ($y = 78.733x + 23.801$, $R^2 = 0.3121$), [c]242 nm ($y = 72.818x + 33.773$, $R^2 = 0.1701$), and [d]279 nm ($y = 57.525x + 16.522$, $R^2 = 0.6155$).

Conclusions

Comparison of univariate models with multivariate (full-spectrum or four-wavelength) models constructed in this work demonstrated that multivariate models provide superior prediction accuracy. This may be due to the fact that overall hydrolysate inhibition is influenced by compounds that absorb in different regions of the UV spectrum, specifically 201, 223, 242, and 279 nm. Therefore, a univariate approach utilizing single absorbance data at any one of these wavelengths may exclude critical information provided by the other three. It is only when multivariate statistical analyses are applied that the best prediction accuracy is achieved. Future work should be aimed at constructing a more comprehensive model by evaluating the effect of increasing calibration-sample size, using different types of lignocellulosic biomass feedstocks, and employing hydrolysates obtained from different pretreatment methods. We note that increasing the size of the calibration set will introduce more variability and unknown interferences into the model. As a result, the model will be more representative of all future samples, which may further improve prediction accuracy.

CHAPTER THREE

Chemometric Modeling of HPLC-UV Data for Identification of Biomass Fermentation Inhibitors

Introduction

Production of fuel ethanol from lignocellulosic biomass holds a broad range of benefits, both environmentally and economically.[9,10] One of the key barriers to cost-effective production of bioethanol has been the lack of a firm understanding regarding process fundamentals, specifically, identification of inhibitory compounds produced during pretreatment. The mechanism of action of such inhibitors also remains largely unknown. Biomass hydrolysates contain many constituents, any of which may be inhibitory to downstream fermentation processes.[28,31,98] Thus, identification of inhibitory hydrolysate constituents is a crucial step in gaining a better understanding of pretreatment and its downstream effects on microbial fermentation processes.

Traditional methods for identification of inhibitory compounds generally involve addition of one or more known degradation products to fermentation media and observation of the resulting effect on batch fermentation of ethanol.[40,45,90] These bottom-up toxicity tests are extremely time-consuming and labor-intensive. Furthermore, protocols used in various laboratories are inconsistent, measuring different output variables and using different fermentation procedures, inhibitor concentrations and/or fermentative organisms. Therefore, comparing results between studies is difficult. Given that hundreds of components are present in pretreatment hydrolysates, many of which are not fully characterized, identification of inhibitory compounds by traditional means is

unlikely to occur on a timescale that is consistent with current mandates for commercial production of cellulosic ethanol in the United States (Energy Independence and Security Act of 2007).[93] Consequently, there remains a need to develop a more rapid and accurate method for identifying hydrolysate constituents responsible for observed inhibitory effects.

Our laboratory is actively investigating a paradigm-shifting approach for identification of inhibitory compounds in hydrolysates based on chemometric modeling of chromatographic data. Chromatographic analyses provide compositional information about the hydrolysate, while fermentation analyses provide information regarding overall toxicity. Chemometric tools minimize co-linearity in the data while providing important information regarding correlations between independent and dependent variables.[74,80,105] Most chemometric analyses of chromatographic data fit in to one of three categories: 1) qualitative analyses, 2) quantitative analyses, and 3) optimization studies. Qualitative analyses generally involve the use of principal component analysis (PCA), soft independent modeling of class analogy (SIMCA), or other statistical methods for classification of measured data. For example, Lonni and coworkers demonstrated that SIMCA may be successfully used to build a chemometric model for classification of three species of *Baccharis* plant, using their HPLC-DAD chromatographic data.[84] Principal component analysis was applied to chromatographic data of 74 alcoholic extracts of three *Baccharis* species to construct three models, one for each species class. Successful classification of eight independent samples to each of the three classes was achieved using the developed models.

Quantitative analyses involve the construction of a predictive model by correlating the independent data set with the concentration of an analyte of interest. For example, in a study by Donmez and coworkers, the use of PLS modeling of HPLC-DAD chromatographic data was investigated for the simultaneous detection of four pharmaceutical compounds in cough syrup.[86] Twenty five synthetic syrup samples were prepared using various concentrations of potassium guaiacolsulfonate (PG), guaifenesin (GU), diphenhydramine HCl (DP), and carbetapentane citrate (CP). A chemometric model was constructed, correlating HPLC-DAD chromatographic data of the synthetic syrup samples with known concentrations of PG, GU, DP, and CP, using PLS-2. The developed model was successfully used to predict concentrations of the four compounds in a commercial syrup sample. Chemometric analyses of HPLC-DAD chromatographic data have also been successfully used for analysis of isoflavones in *Trifolium lucanicum* plant and pesticides in water samples.[107,108]

Optimization studies often involve the use of response surface methodology (RSM) to build a second-order polynomial model, which can be used to optimize a variable of interest that is influenced by several independent variables.[109] For example, in a study by Sivakumar and coworkers, RSM and Derringer's desirability function were used to optimize an HPLC method for the separation of two pharmaceutical drugs, Domperidone (DP) and Pantoprazole (PP).[110] Independent variables selected for optimization included acetonitrile concentration in the mobile phase, phosphate buffer molarity and flow rate. Response factors selected for optimization included the retention factor of PP, retention times of PP, DP, and the internal standard (IS) acetophenone, and the resolution between two pairs (PP-IS and IS-DP). A central composite design (CCD)

was used to fit all six variables into a second-order polynomial to determine the optimum conditions for separation of the studied compounds. The predicted optimized HPLC conditions were subsequently applied for effective separation of both compounds in real pharmaceutical tablets.

While numerous studies have reported the use of chemometrics for qualitative, quantitative, and optimization analyses, the use of chemometrics for qualitative identification of specific variables within chromatographic data has not been reported. In this work, we report a novel method for identification of inhibitory hydrolysate constituents using chemometric modeling of chromatographic data. Partial least squares can be used to correlate chromatographic data of raw hydrolysates with microbial inhibition. The model can then be used to identify specific variables (i.e. retention times) that have the strongest correlation with inhibition. Once identified, these retention times can be further investigated to reveal the identity of compounds that are likely to be most inhibitory to the test organism used in fermentation of hydrolysates. While most studies utilize chemometrics to find correlations that can be used for prediction purposes (qualitative or quantitative), the approach presented here utilizes chemometric modeling for qualitative identification of specific variables that significantly contribute to the observed effect (i. e. inhibition). The proposed methodology is a more rapid, top-down approach for identification of inhibitory hydrolysate constituents, in that it utilizes compositional information regarding whole hydrolysate, taking into account additive and/or synergistic effects from all constituents which may affect toxicity. Furthermore, the approach does not require prior knowledge of all hydrolysate constituents, given that compositional information is obtained via qualitative analyses.

Herein we provide proof-of-concept data demonstrating the feasibility of our approach. Twenty-six dilute-acid hydrolysate samples of varying composition were qualitatively characterized using reversed-phase liquid chromatography with UV detection (RPLC-UV) and ion chromatography with conductivity detection (IC). Microbial inhibition was assessed independently in batch fermentation experiments. Chemometric tools were used to find useful correlations between inhibition and chromatographic data. The quality of chemometric models was evaluated by comparing predicted inhibition of five samples that were kept out of the calibration sample set with experimental values. The x-loadings feature was used to identify retention times that were most strongly correlated with microbial inhibition. Compounds eluting at these retention times were known inhibitors of fermentative microorganisms.

Experimental: Materials and Methods

Reagents and Standards

Analytical grade glucose, yeast extract, peptone, ethanol, sulfuric acid and citric acid monohydrate were purchased from Sigma-Aldrich (St. Louis, MO, USA). Five ethanol standards at concentrations of 10, 4, 2, 1, and 0. 2 mL/L were prepared from an ethanol (100 mL/L) stock solution. These standards were subsequently used to build an external standard calibration curve for high performance liquid chromatography (HPLC) analyses of ethanol concentration. For IC analyses, a stock standard consisting of 42 mg/L acetic acid, 50 mg/L propionic acid, and 254 mg/L sodium chloride in deionized water was used to prepare calibration standards via serial dilution. Distilled water was

purified and deionized to 18 MΩ with a Barnstead Nanopure Diamond UV water purification system.

Preparation of Hydrolysates and Subsequent Fermentations

Hydrolysate samples were supplied by Dr. Peter van Walsum (University of Maine, Forest Bioproducts Research Initiative), and fermented previously.[111] Once hydrolysates were received, the pH was adjusted to 4.85 using calcium hydroxide. Particulates were removed by filtration. All hydrolysate samples were subsequently batch fermented in triplicate experiments. Samples were drawn prior to inoculation and from time 0 to 48 hours following inoculation. Samples were subsequently analyzed via HPLC with refractive index detection to determine ethanol concentration.[111] Table 3.1 lists observed inhibition values for the 26 hydrolysates and the corresponding pretreatment conditions used to generate each sample.

Reversed-Phase Liquid Chromatography Analysis of Raw Hydrolysates

Reversed-phase LC analyses generally followed the procedure reported by Chen et al.[30] However, samples did not undergo MTBE extraction prior to analysis. This step was omitted in the present study to preserve the compositional integrity of hydrolysates. Filtered hydrolysates were diluted 40-fold with deionized water and analyzed directly with UV detection at four wavelengths. Additionally, a modified mobile-phase gradient consisting of (A) 0.025% (v/v) formic acid in water and (B) 10:90 water-acetonitrile containing 0.025% (v/v) formic acid was employed to achieve chromatographic separation.[31] UV absorbance was monitored at 201, 223, 242, and 279 nm. These wavelengths were identified by a previous chemometric model utilizing UV-visible

spectral data (190-450 nm), as being most strongly correlated with microbial inhibition.[111]

Table 3.1. Pretreatment severity and observed percent inhibition for calibration and validation samples

Sample number[a]	Time (min)	Temperature (°C)	CSF[b]	Percent inhibition at 14 hours[c]
1	16	190	2.70	63
2	64	170	2.72	41
3	24	185	2.73	38
4	50	175	2.76	38
*5	14	195	2.79	48
6	40	180	2.81	0
*7	60	175	2.84	70
8	22	190	2.84	46
*9	64	175	2.86	25
10	12	200	2.87	39
*11	35	185	2.90	50
12	26	190	2.91	46
13	55	180	2.95	36
14	20	195	2.95	57
15	60	180	2.98	70
16	22	195	2.99	71
17	16	200	3.00	50
18	24	195	3.03	38
*19	18	200	3.05	60
20	55	185	3.09	68
21	40	190	3.10	62
22	30	195	3.12	44
23	24	200	3.17	54
24	50	190	3.20	42
25	45	195	3.30	57
26	64	190	3.31	58

[a] Validation samples are represented by asterisk

[b] Corn stover hydrolysates were generated using ASE 300, 0.7% (w/w) H_2SO_4, 8% (w/v) solids concentration

CSF = log [t x exp ($T_r - T_b$)$/$14.75] – pH

[c] Mean percent inhibition from triplicate fermentations.

Ion Chromatography Analysis of Raw Hydrolysates

Ion chromatography analyses were carried out using a previously reported procedure.[16] Briefly, 250-µL of filtered hydrolysate was loaded onto a preconditioned Supelclean LC−18 SPE cartridge. The cartridge was rinsed with slightly less than 5 mL of deionized water and the eluate was diluted to a final volume of 5 mL prior to analysis. An aqueous NaOH step-gradient was employed to achieve chromatographic separation: 1 mM, 0.0-20.0 min; 30 mM, 20.0-21.0 min; 60 mM, 21.0–22.0 min; and back to 1 mM, 22.0–27.0 min. Eluted sample components were monitored via conductivity.

Chemometric Analysis

Unscrambler 9.8 (CAMO Software Inc., Woodbridge, NJ, USA) was used for chemometric analyses. Chromatographic data and mean (n=3) inhibition values were imported into the model as input variables and mean-centered prior to all chemometric analyses. Tri-PLS−1 was used for regression of HPLC-UV data, given that chromatographic data were obtained in a three-dimensional format, with the axes corresponding to retention time, wavelength, and absorbance. Partial least squares (PLS−1) was used for regression of IC data. All chromatographic data included in chemometric models were derived from a single injection of each sample. The method used to calibrate the models was leave-one-out cross-validation.

Five hydrolysate samples that were not included in the calibration sample set were selected for validation purposes. Mean inhibition values of all 26 samples were placed into five categories, with each category corresponding to a specific inhibition range (20-30%, 30-40%, 40-50%, 50-60%, and 60-70%). Validation samples (5, 7, 9, 11, and 19) were randomly selected from each of the five categories, using a random number

generator. This process ensured that inhibition values of the validation samples spanned the range of inhibition values used for model calibration,

Results and Discussion

HPLC-UV Chromatographic Data

Figures 3.1-3.4 depict HPLC-UV chromatograms of all 26 hydrolysate samples, obtained at 201, 223, 242, and 279 nm. Based on previous work in our laboratory, it is known that analyte retention is primarily governed by hydrophobic interactions between the column and the analyte.[30] Generally, low-molecular weight aliphatic acids elute first, followed by aldehydes, phenols, and aromatic acids. The most prominent peaks in each chromatogram occurred at 14.6 and 16.2 minutes. As depicted in Figures 3.1-3.4, hydrolysates have many constituents, any of which can be inhibitory to fermentation. Given the complex nature of the HPLC-UV chromatograms, identification of specific retention times whose change in peak area (i.e. concentration) is most strongly correlated with inhibition would be difficult at best by visual comparison.

Chemometric Model Construction and Validation

Reversed-phase LC data at all four monitored wavelengths and inhibition data for the 21 calibration samples were imported into Unscrambler as input variables. Tri-PLS algorithm was used to develop the initial regression model. Partial least squares regression focuses on variance in the independent data set that is correlated with the depend dependent variable while minimizing the influence of variance that is uncorrelated with the dependent variable.[72] The model was calibrated using leave-one-

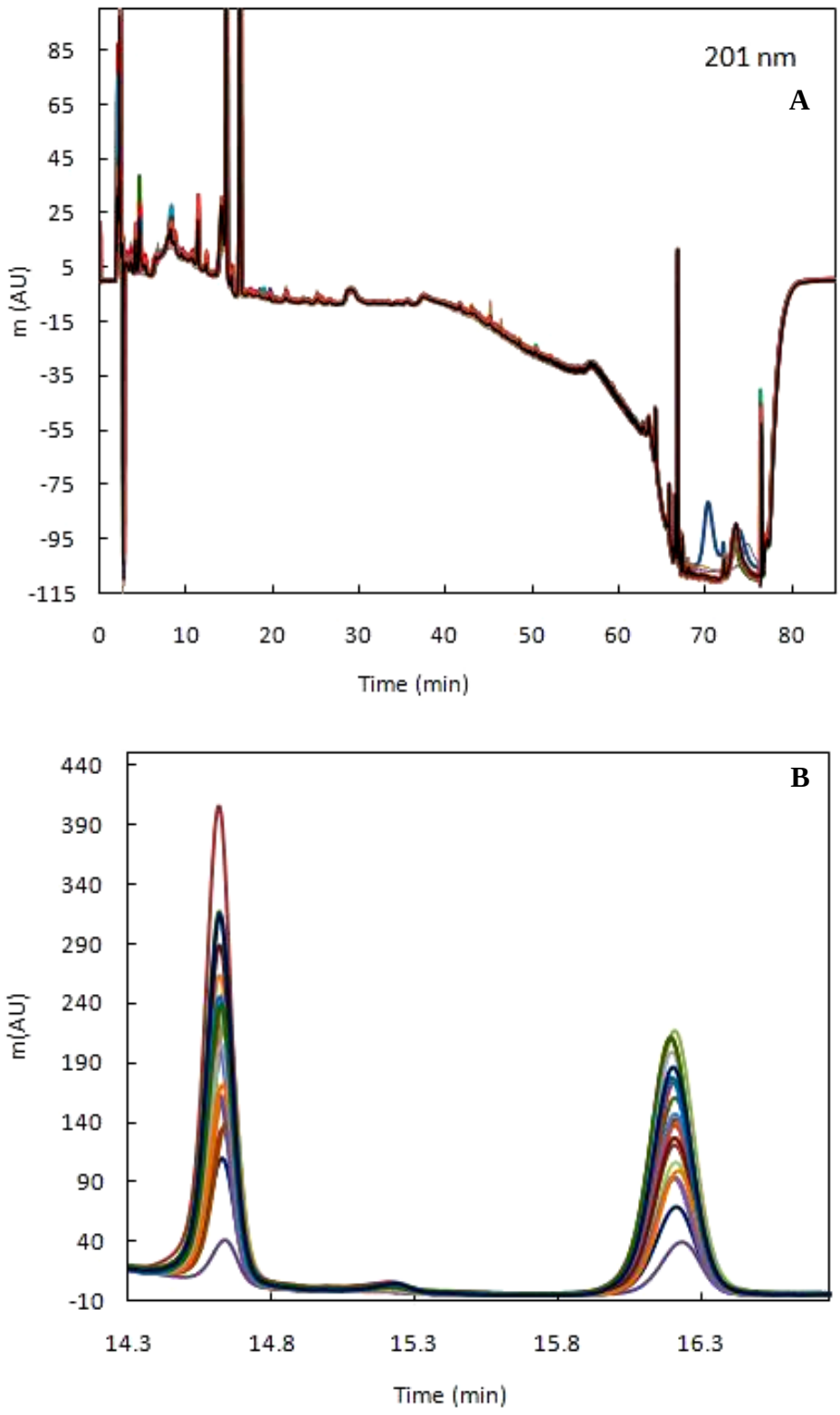


Figure 3.1. HPLC-UV chromatograms of raw hydrolysates. A) Chromatograms at 201 nm, B) Expansion of chromatograms shown in A in the 14.3-16.75 min region.

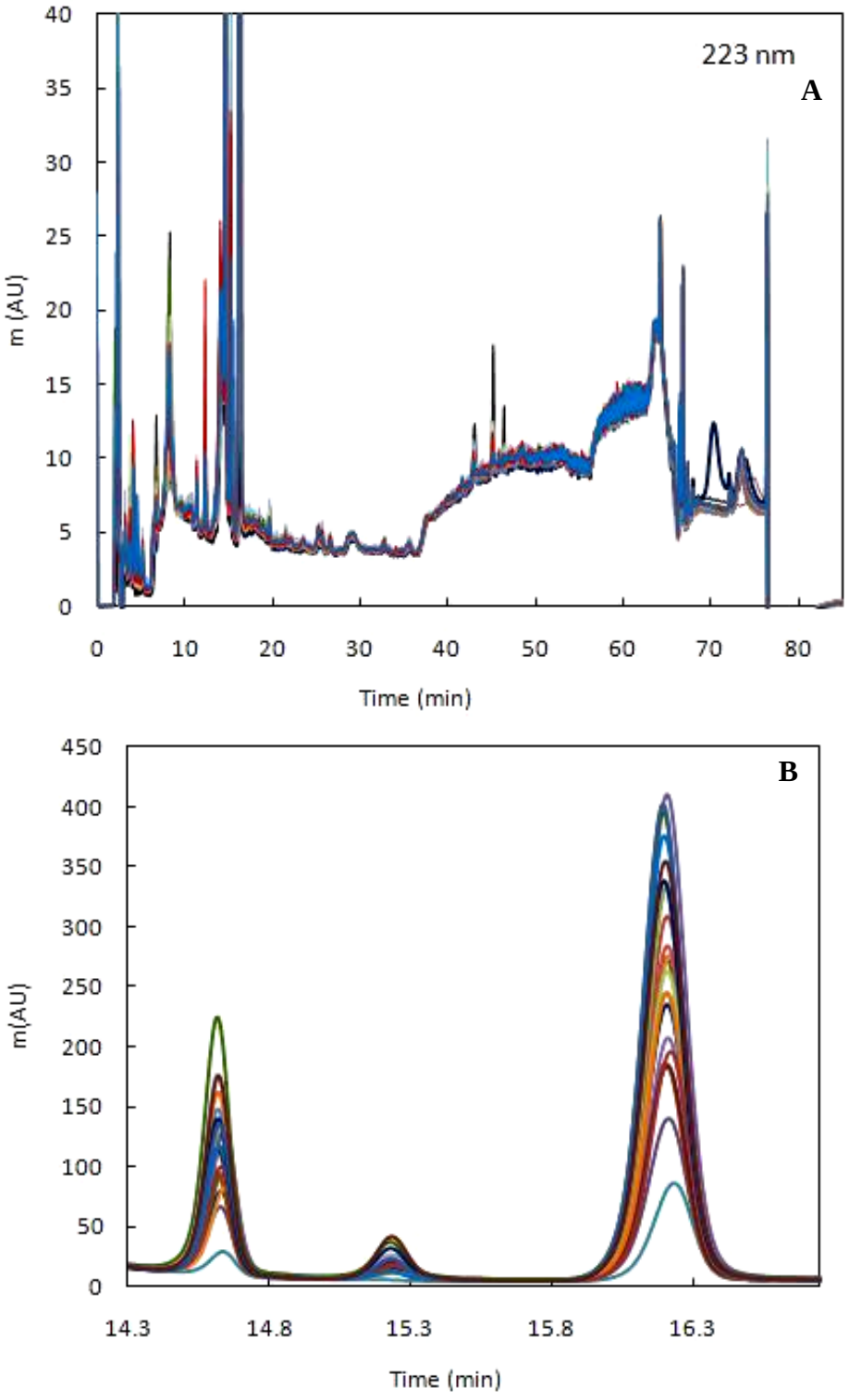


Figure 3.2. HPLC-UV chromatograms of raw hydrolysates. A) Chromatograms at 223 nm, B) Expansion of chromatograms shown in A in the 14.3-16.75 min region.

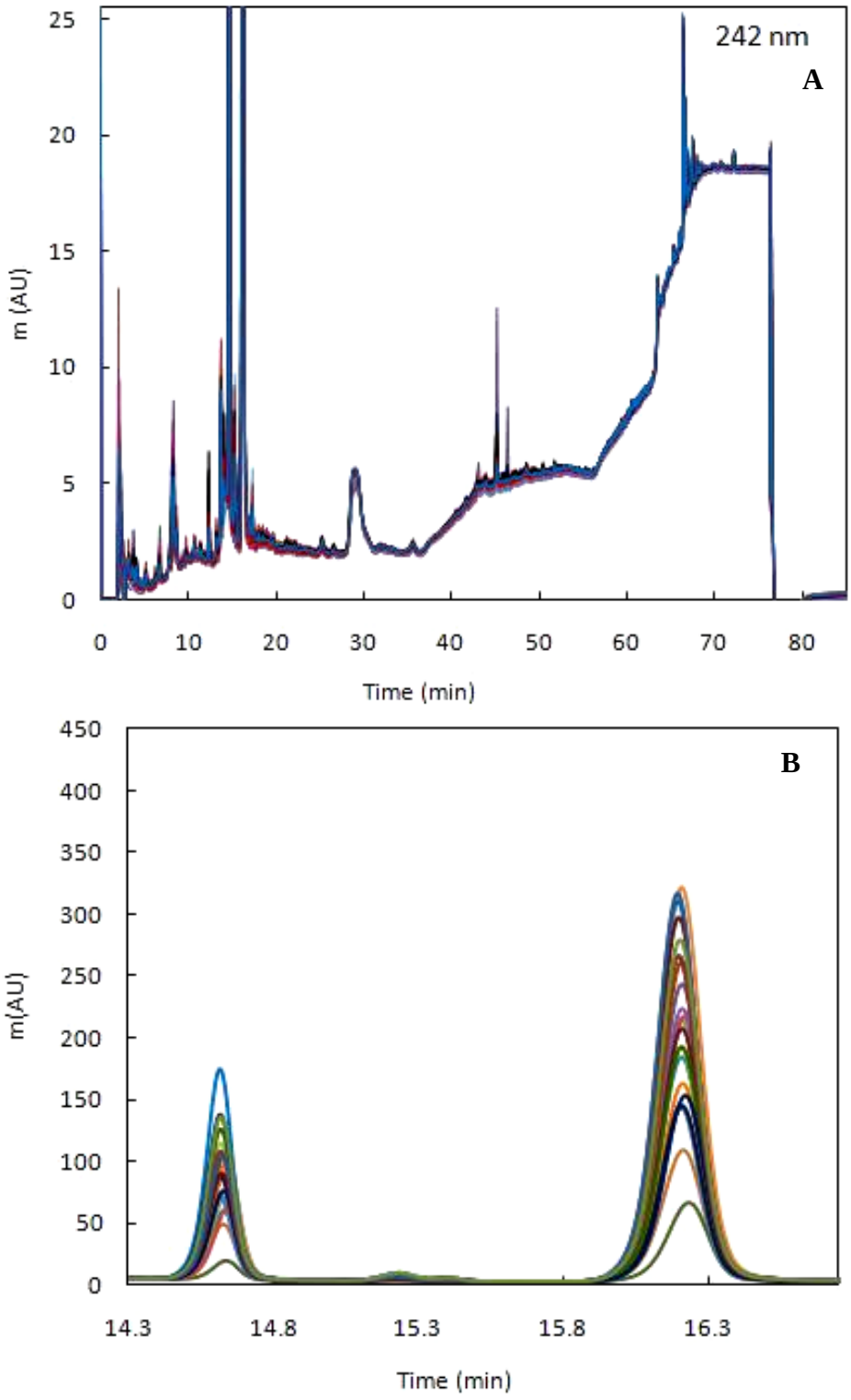


Figure 3.3. HPLC-UV chromatograms of raw hydrolysates. A) Chromatograms at 242 nm, B) Expansion of chromatograms shown in A in the 14.3-16.75 min region.

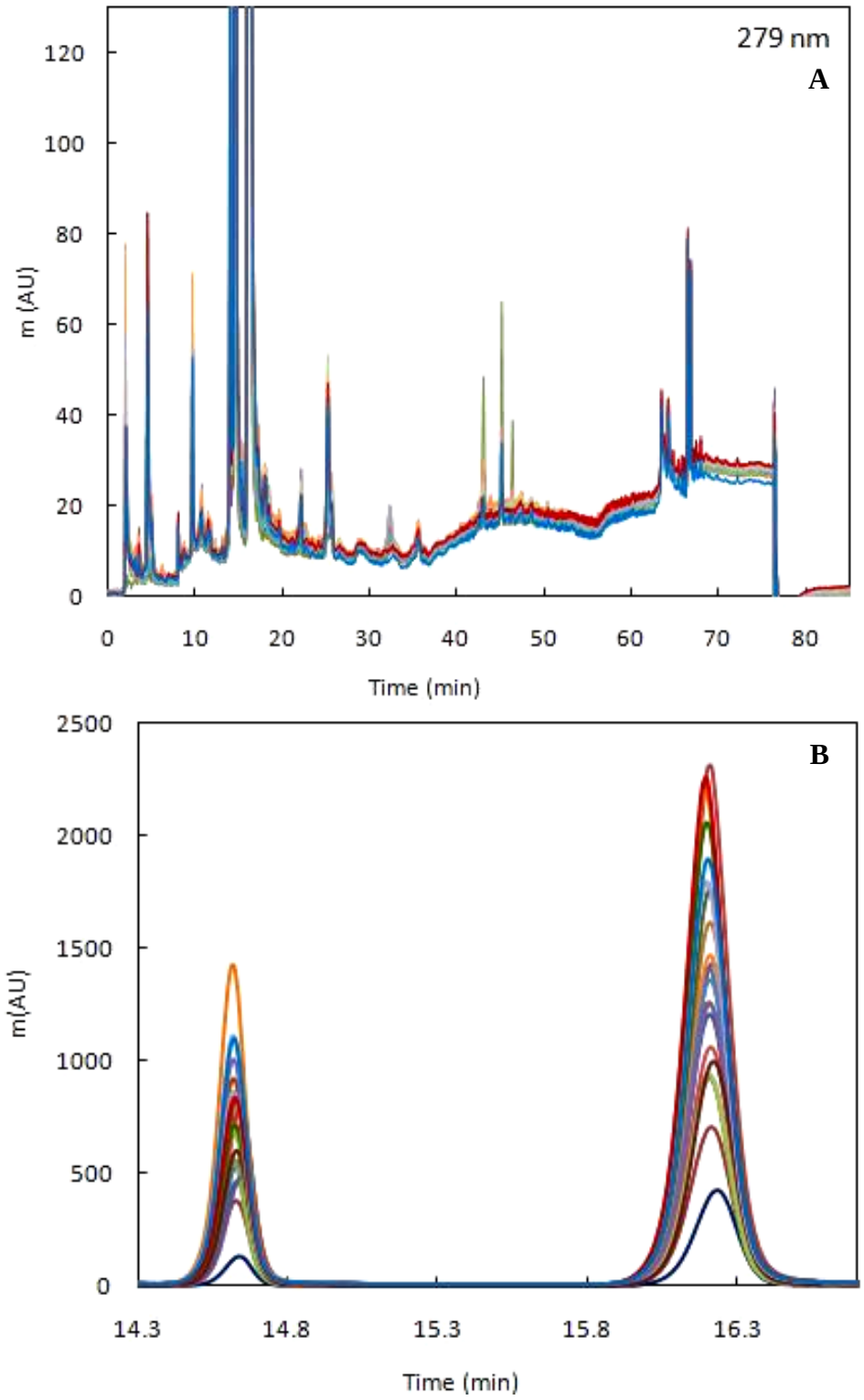


Figure 3.4. HPLC-UV chromatograms of raw hydrolysates. A) Chromatograms at 279 nm, B) Expansion of chromatograms shown in A in the 14.3-16.75 min region.

out cross-validation. during which, the model was repeatedly recomputed while one calibration sample was left out and subsequently predicted using the developed model.[73,80] This process was repeated until every sample had been left out once. In each case, the difference between predicted values and experimentally determined values of all samples were used to calculate the predicted residual error sum of squares (PRESS).

Figure 3.5A shows the plot of residual validation variance as a function of number of principal components (PCs) used to construct the model. Residual variance was calculated by averaging PRESS (obtained during leave-one-out cross validation) over all calibration samples (n=21).[72] This was repeated for models using a different number of PCs. The residual validation variance plot represents the error that is expected with future predictions, depending on the number of PCs used. Figure 3.5A suggests that using models with 1-3 PCs would yield a lower residual variance than models constructed with a larger number of PCs. To determine the optimum number of PCs, models constructed with 1-3 PCs were used to predict inhibition of five hydrolysate samples that were not used to construct the model. The corresponding root mean square error of prediction (RMSEP) for models using 1, 2, and 3 PCs were 6, 5, and 7, respectively, suggesting that the best prediction model was constructed with 2 PCs.

Figure 3.5B shows the scores plot, which displays the location of samples in the coordinate system for the first two PCs. It is demonstrated below the x-axis that the first two PCs collectively explain 99% of the variance in the chromatographic data and 45% of the variance in the inhibition data, respectively. In a perfect model, samples would be uniformly distributed throughout the coordinate system. For the model developed here, most samples were located near the center and right regions of the coordinate system.

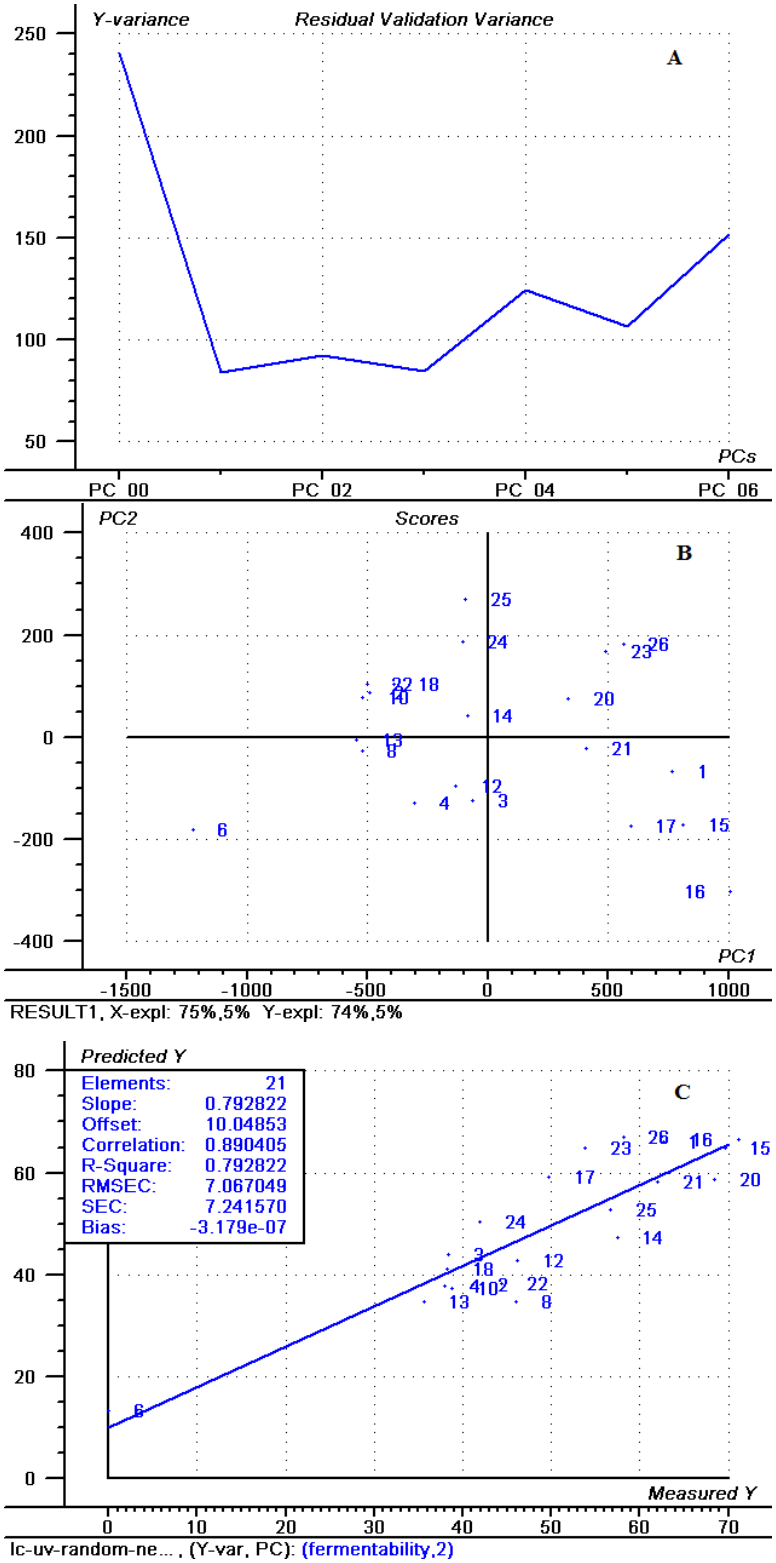


Figure 3.5. Regression overview for model constructed with RPLC-UV data: A) variance vs. number of components, B) scores plot, C) predicted vs. measured plot

Although a more uniform distribution would likely result in a more robust model, the data swarm was deemed acceptable for the purposes of this work.

Once calibrated, the regression vector was used to predict inhibition for every calibration sample. The predicted values were then plotted against measured values (Figure 3.5C) to evaluate the model's fit (i.e. the degree of correlation between chromatographic data and observed inhibition). The more linear the regression vector (i. e. the higher the R^2), the better the correlation would be. A perfect model would have an R^2 of 1. For the developed model, R^2 was 0.79. The offset (intercept) was 10. RMSEC, standard error of calibration (SEC), and bias (systematic error) were 7, 7, and -3 E^{-7}, respectively, indicating negligible systematic error and reasonable accuracy. The correlation coefficient is the square-root of R^2 and is a measure of the correlation between HPLC-UV chromatographic data and observed inhibition of calibration samples. This value can range from −1 to 1, with 0 indicating no correlation. The sign of the coefficient indicates a negative or positive correlation. In Figure 3.5C, the correlation coefficient is 0.89, indicating a positive correlation between chromatographic data and inhibition (i.e. an increase in inhibition is expected with an increase in absorbance). Although the correlation coefficient and R^2 can be used to evaluate the model's correlation, the best way to assess its quality is to validate it using new samples.

Table 3.2 shows the results of the validation step, including experimental and predicted inhibition for each sample (columns two and three) along with RMSEP. Given that all hydrolysate samples were fermented in triplicate under similar conditions, pooled standard deviation (s_{pool}) was used to determine experimental precision. Four of five samples (5, 7, 9, and 11) had predicted inhibition values that were within 95% confidence

limits of experimental determinations. These results indicate a reasonably strong correlation between microbial inhibition and reversed-phase chromatographic data.

Table 3.2. Experimentally determined inhibition and predicted inhibition for validation samples.

Validation sample number	Experimental % inhibition[a]	Predicted-model 1[b]	Predicted-model 2[c]	Predicted-model 4[d]
5	48 ± 6	48	51	51
7	70 ± 6	75	80	59
9	25 ± 6	28	29	26
11	50 ± 6	52	47	37
19	60 ± 6	71	62	56
RMSEP		5	5	8

a mean fermentability from triplicate fermentations of hydrolysate and corresponding 95% confidence intervals. Pooled standard deviation was 5.
b model constructed using RPLC-UV chromatograms from all four wavelengths
c model constructed with hydrolysates spiked with furfuryl alcohol
d model constructed with chromatograms obtained from ion chromatography analyses

It is important to note that the above predictions were carried out using the developed model to evaluate its correlation and not for assessing its use as a predictive model. A predictive chemometric model has been previously constructed in our laboratory, correlating UV-visible spectral data of hydrolysate samples with microbial inhibition. This model is better suited as a predictive model compared to the model developed in this study, given that UV-visible analyses are less time-consuming and labor-intensive than chromatographic analyses. In the present work, the purpose of developing the LC-UV model was to identify retention times that were most strongly correlated with inhibition.

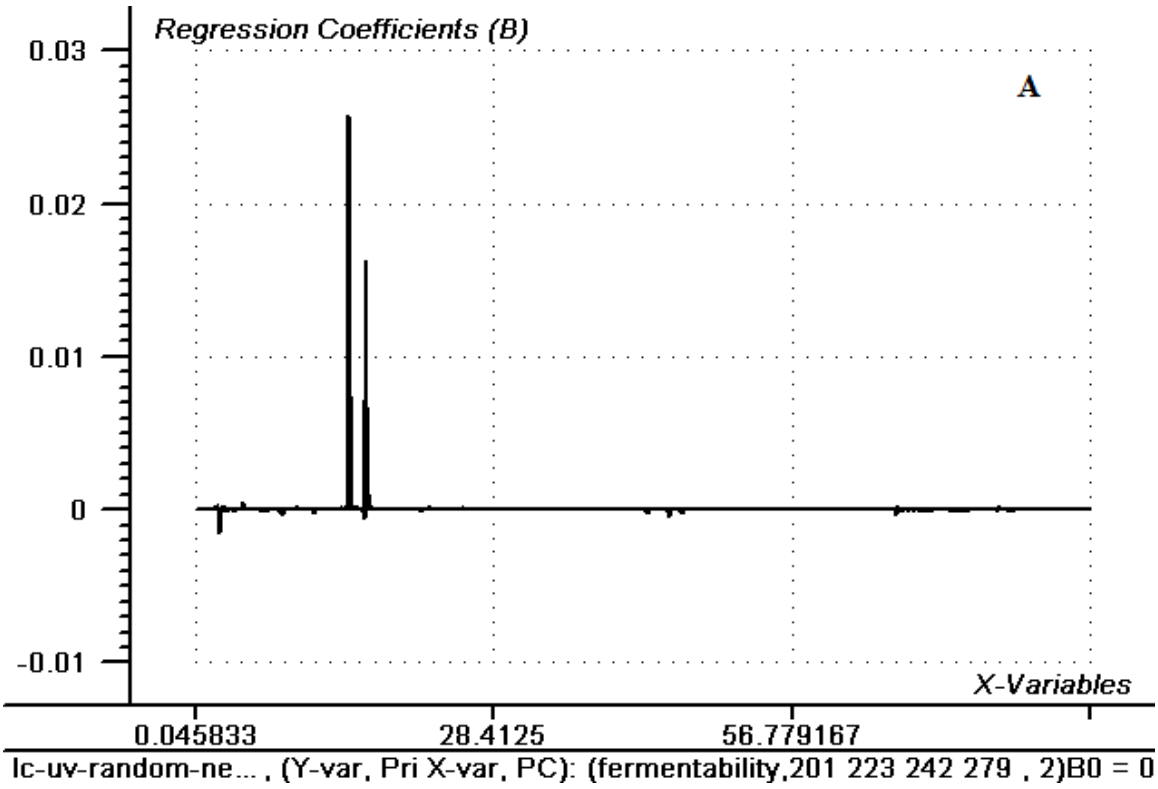
Identification of Significant Retention Times

Once the model was successfully validated, two chemometric tools (regression coefficient plot and x-loadings) were used to identify the significant variables. The constructed model was explained by a regression vector in the form of

$$\hat{y} = b_o + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n \quad (3.1)$$

where $\hat{y}$ is inhibition for a given sample, b is the regression coefficient determined by the algorithm, and X_n is the measured absorbance at a given wavelength and retention time. The regression coefficients (b) are plotted in Figure 3.6A as a function of retention time. Variables with larger regression coefficients make a more significant contribution to the correlation between chromatographic data and inhibition. Peak maxima in Figure 3.6A are at 14.6 and 16.2 minutes, suggesting that a majority of the difference in sample composition that is correlated with inhibition is due to the change in absorbance (i.e. concentration) of compounds eluting at these retention times.

The loading plot is another tool that can be used along with the regression coefficient plot to identify variables with high significance to the model. The loading is the cosine of the angle between each variable and the corresponding PC.[77] As the angle between the component and the variable approaches 0 or 180 degrees, the loading approaches 1 or −1. This indicates that the component is parallel with the variable, and the variable contributes much information to the component. As the angle between the variable and the component approaches 90 degrees, the loading approaches 0, indicating that little information is contributed to the component from that variable. The higher the loading value, the more significant that variable is to the corresponding PC. Figure 3.6B shows the loadings plot for PC1 and PC2. The most significant retention times in Figure



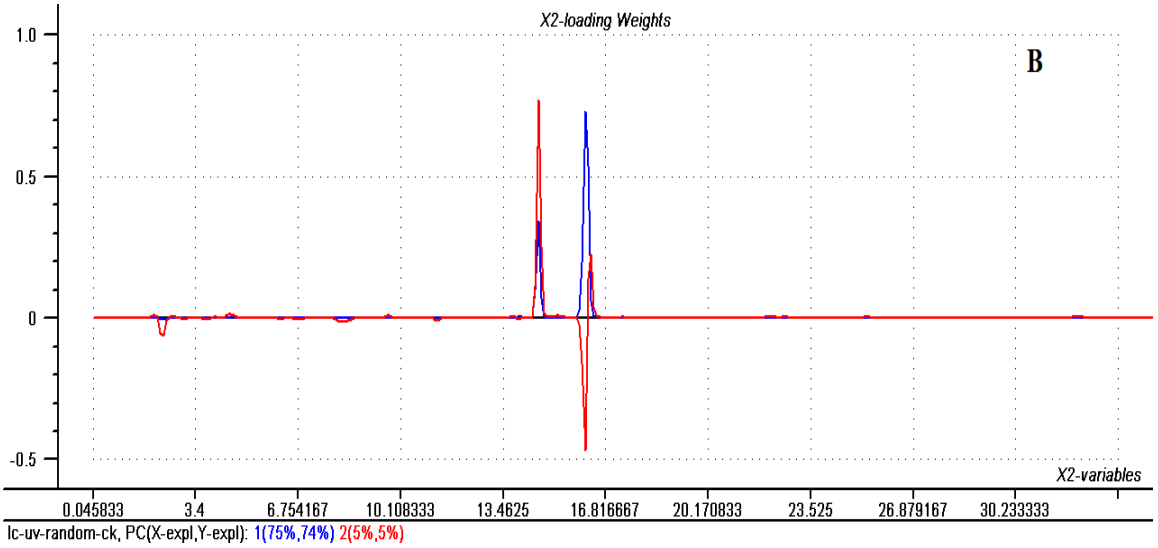


Figure 3.6. A) Regression coefficient plot for HPLC-UV model, B) X-loadings for the first two principal components.

3.6B are identical to those identified by the regression coefficient plot, further confirming their significance to the observed correlation.

Significant retention times identified in Figure 3.6 correspond to 5-hydroxy-2-methylfurfural (HMF) and furfural, respectively, which are of particular importance for toxicity in yeast.[33,37] Identification of both compounds was confirmed by comparing UV absorbance ratios at multiple wavelengths ($A_{\lambda 1}/A_{\lambda 2}$) and retention time data with a reference standard (90%> agreement). Additionally, peak purity for both compounds was evaluated by comparing multiple absorbance ratios at several retention times within the peak window with a reference standard (90%> agreement). Furfural has shown to inhibit ethanol production by 31% at concentrations as low as 1 g/L, when tested alone.[40] At 2 g/L, furfural further inhibited cell growth of *S. cerevisiae*. HMF has shown to inhibit ethanol production by 20% when tested alone at 2 g/L. When tested together, a total furan concentration of 3 g/L was required to inhibit ethanol production.[33] However, the minimum inhibitory concentration of furans is significantly lower when they are present in a mixture of other degradation products, which may be due to synergistic effects. For example, Martinez and coworkers demonstrated that when the ratio of HMF to furfural is low (0.15-0.2), inhibition of ethanol production occurred when the total furan concentration exceeded 0.2 g/L.[42] Even as the ratio of HMF to furfural increased to 0.8−1.1, the minimum inhibitory concentration of furans was as low as 0.5 g/L. Several other studies have also demonstrated the inhibitory effects of HMF and furfural.[34,41,44] The concentration range for HMF and furfural observed in samples used in this study was 0.04-0.4 and 0.1-0.6 g/L, respectively. Accordingly, identification of these compounds

by the model developed here was a promising result, confirming that the methodology employed is effective for identifying inhibitory compounds.

Control Experiments Demonstrating that the Model isn't Simply Picking Prominent Peaks

Given that the retention times identified by the model correspond to the most prominent peaks in the chromatogram, control experiments were carried out to ensure that the model did not simply select the most prominent variables as the most significant ones. To accomplish this, the 21 hydrolysate samples that were used to construct the model were spiked with furfuryl alcohol. Furfuryl alcohol was chosen due to its high absorptivity (ε_{223nm} = 1.3 * 10^3 M^{-1} cm^{-1}) in the UV region and low toxicity in yeast.[37] The concentrations of furfuryl alcohol used in this work ranged from 0.0418 to 0.125 g/L. This concentration range was selected to obtain a similar absorbance range for the furfuryl alcohol peak as the peaks corresponding to 5-HMF and furfural (Figure 3.1B). To ensure that the amount used for spiking did not affect percent inhibition of hydrolysates, three hydrolysate samples with low, mid, and high inhibition (samples 8, 9, and 20) were spiked with furfuryl alcohol at a concentration of 0.125 g/L and re-fermented. A non-spiked version of each sample was also fermented for comparison. Fourteen hours following inoculation, percent inhibition was 46, 25, and 68% for samples 8, 9 and 20, respectively. For unspiked samples, percent inhibition was 43, 26, and 70%, corresponding to a relative percent difference of 7, 4, and 3%. Results confirmed that observed inhibition in spiked samples was not significantly different than that of the unspiked samples. Subsequently, hydrolysate samples were spiked randomly with varying amounts of furfuryl alcohol, ranging between 0.0418 and 0.125 g/L, and re-

analyzed via RPLC with detection at 201, 223, 242, and 279 nm. Independent experiments confirmed that the furfuryl alcohol spiking levels were within the linear range of the detector (Figure 3.7). Figures 3.8 illustrates HPLC-UV chromatograms corresponding to samples showing minimum and maximum absorbance at 201 and 223 nm.

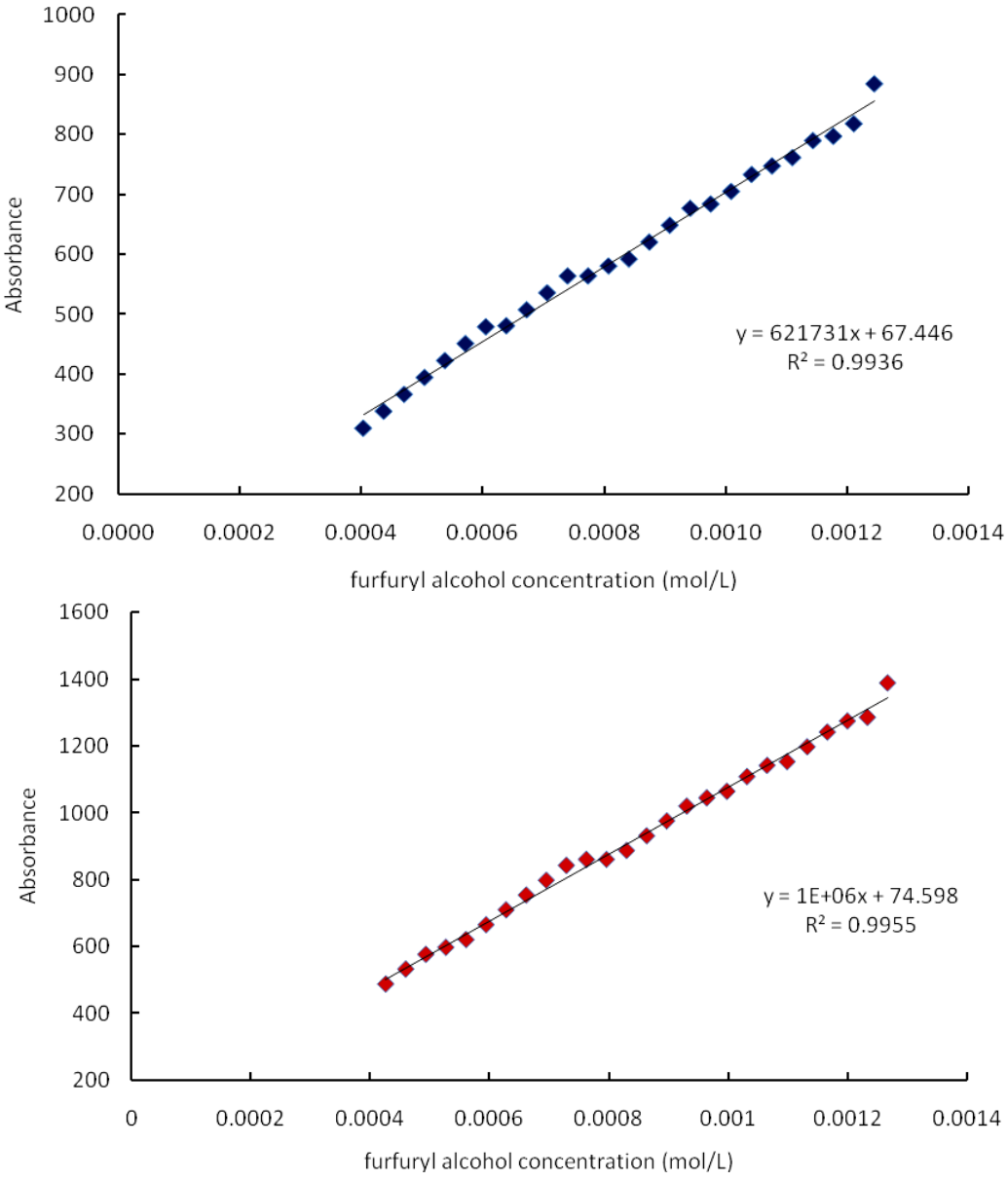


Figure 3.7. Plot of absorbance as a function of concentration for all hydrolysate samples spiked with furfuryl alcohol. Top: Absorbance at 201 nm, Bottom: Absorbance at 223 nm

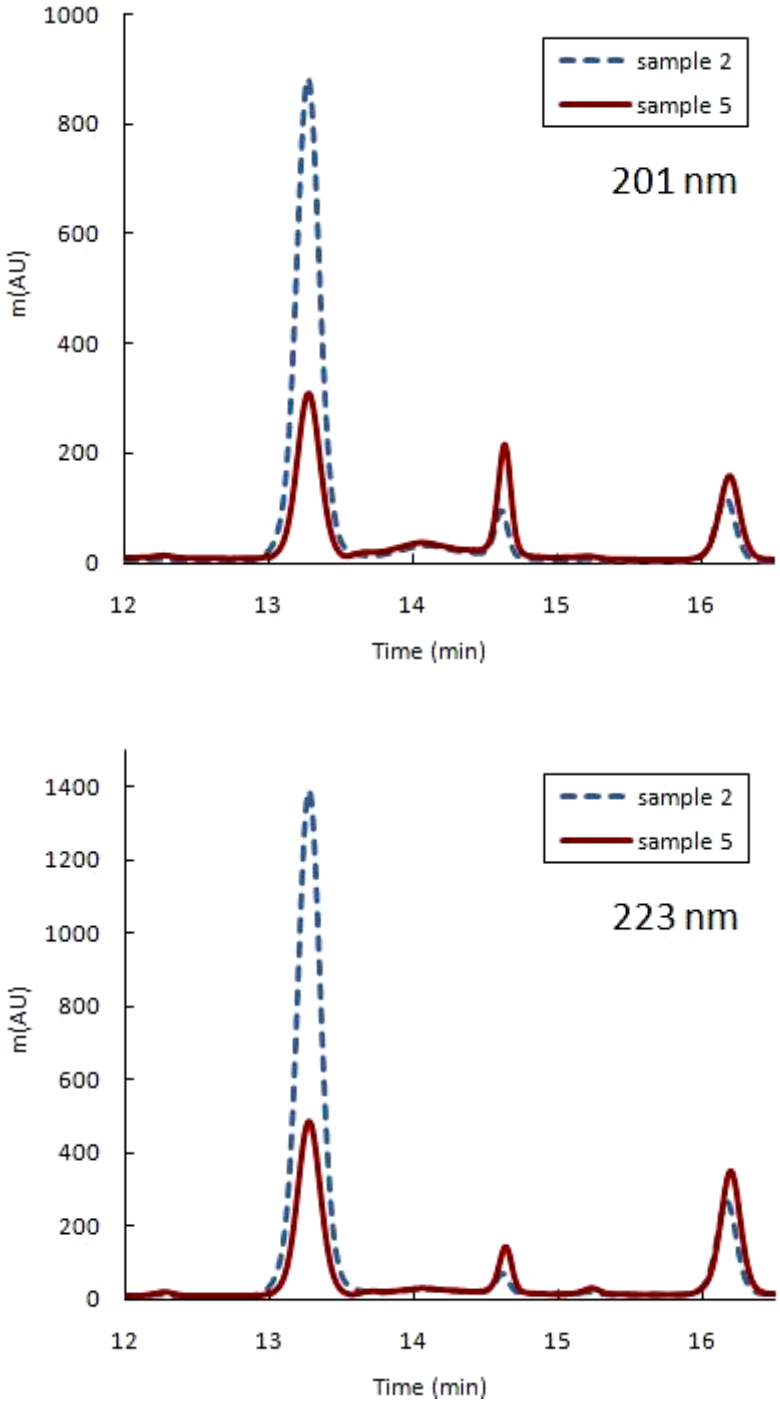


Figure 3.8 HPLC-UV chromatograms of spiked hydrolysate samples at 201 and 223 nm.

Chromatograms of the 21 spiked hydrolysate samples from all four wavelengths and the original inhibition data were used to re-construct the chemometric model. The regression overview for this model is provided in Figure 3.9. Overall, this model had comparable results, using two PCs, to the original model constructed with RPLC-UV data of unspiked hydrolysates. This was anticipated, given that no correlation was expected to exist between concentration changes in furfuryl alcohol and microbial inhibition. This model was used to predict inhibition for the five validation samples that were spiked and analyzed in the same manner as the calibration samples. Prediction results for this model are presented in Table 3.2 (column 4). The model had an R^2 of 0.72. RMSEP was 5, a comparable value to model 1. Furthermore, predicted values were within the 95% confidence intervals of experimentally determined inhibition in four of five validation samples (5, 9, 11, and 19). The retention times identified by the regression coefficient and loadings plots (Figure 3.10) were unchanged relative to model 1, and corresponded to expected retention times for 5-HMF and furfural. This confirms that the chemometric model developed did not identify significant retention times based on peak prominence, but based on changes in peak area and/or height and how the change is correlated with microbial inhibition.

Exclusion of Significant Variables for Identification of Additional Retention Times

Figures 3.1-3.4 depict several peaks in the RPLC-UV chromatogram, with the most prominent ones at 14.6 and 16.2 minutes. Since the prominent peaks were identified by the model as the most significant variables, it was unclear whether the significance of smaller peaks was masked by the presence of prominent ones. To

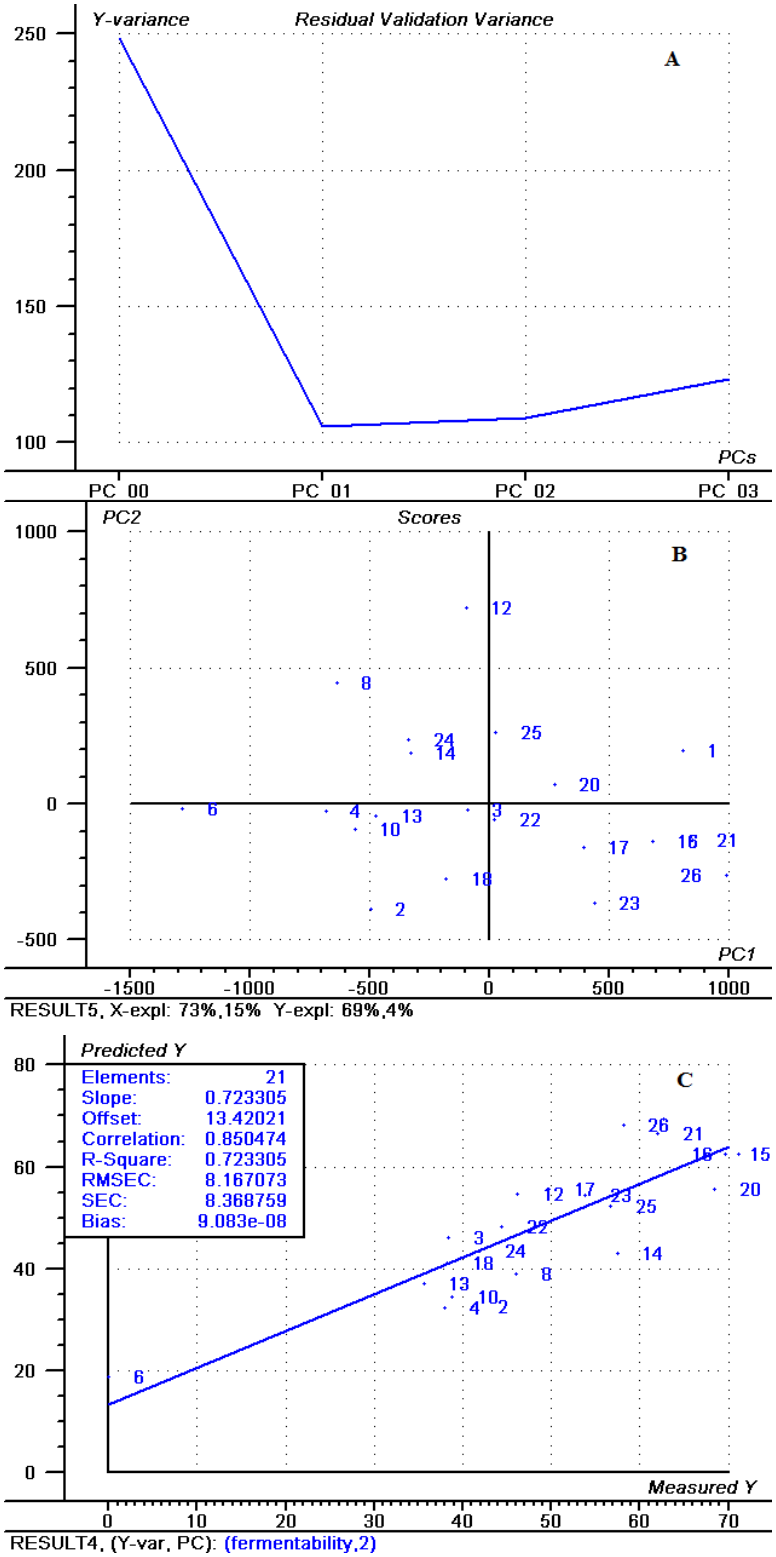


Figure 3.9. Regression overview for model constructed using spiked hydrolysates: A) variance vs. number of components, B) scores plot, C) predicted vs. measured plot

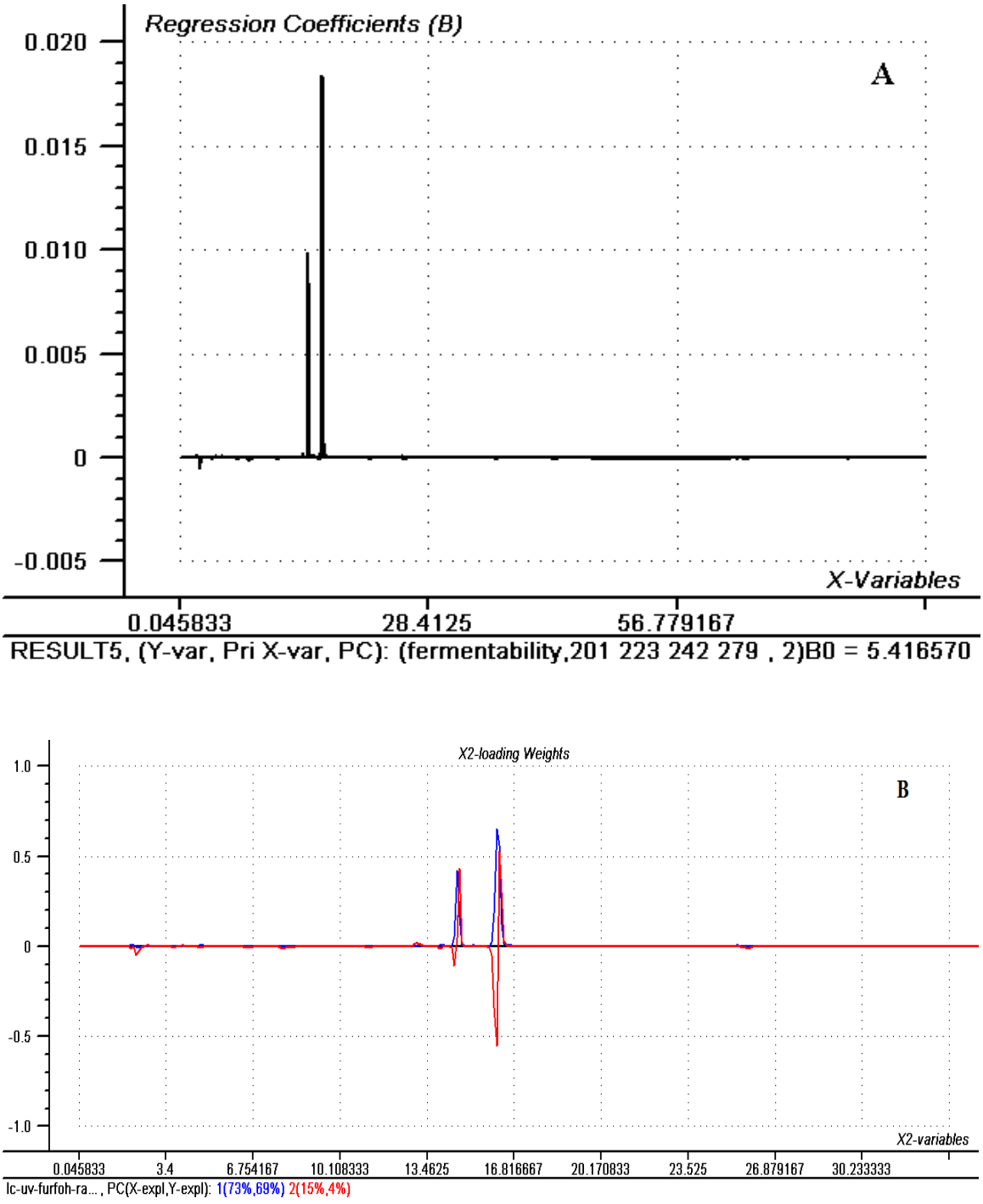


Figure 3.10. A) Regression coefficient plot for model constructed using spiked hydrolysates, B) X-loadings for the first two principal components.

evaluate this, the retention time window ranging from 14−17.5 minutes was excluded from the chromatograms obtained on raw hydrolysates and the model was reconstructed with edited chromatograms. The regression overview for this truncated model is presented in Figure 3.11. Overall, the model's correlation was poor. Results of this experiment demonstrate that utilizing truncated chromatograms did not result in identification of additional retention times that were correlated with inhibition. Results also confirm that the chromatographic data for 5-HMF and furfural made a significant contribution to the initial RPLC-UV model and that the change in peak areas of these compounds had a stronger correlation with inhibition than any other eluting compound.

Use of a Different Detection Method for Identification of Additional Retention Times

While it was promising that the RPLC-UV model identified retention times corresponding to HMF and furfural, it was surprising that the model did not also identify the retention time corresponding to elution of acetic acid as a significant variable. Previous studies have suggested that acetic acid is inhibitory to microbial growth and fermentation.[15,43,92] The inability of the model to identify the acetic acid peak as a significant retention time could be due to the methodology employed. Acetic acid has low absorbance in the UV region, is not strongly retained, and could be co-eluting with other polar sample constituents. Therefore, better resolution and/or more selective detection may lead to identification of the acetic acid peak by the model as a significant variable. To test this hypothesis, all hydrolysate samples used for model construction were analyzed by ion chromatography (IC) with conductivity detection. The gradient separation employed was designed to resolve early eluting compounds observed in

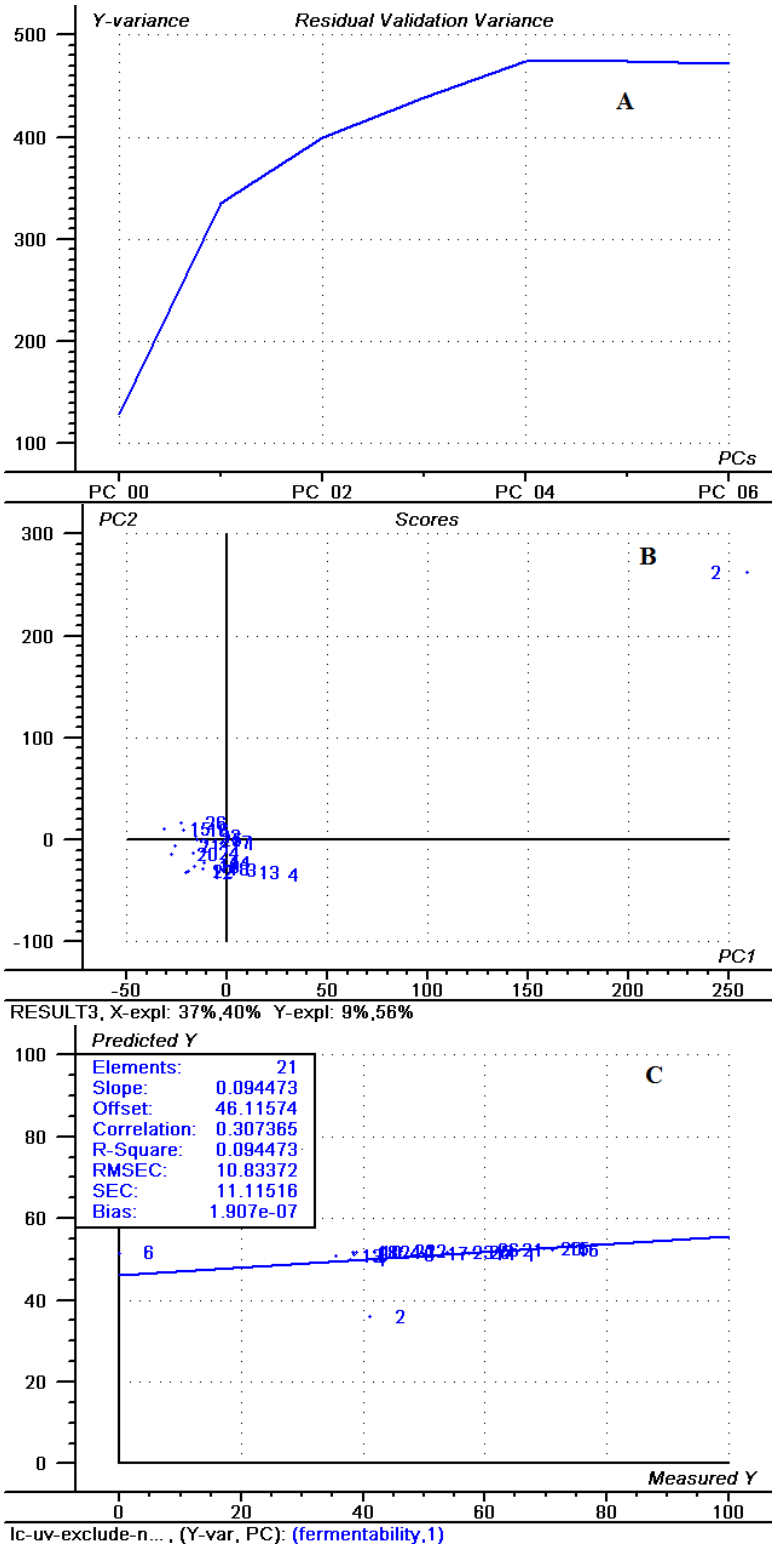


Figure 3.11. Regression overview for model constructed using truncated chromatograms: A) variance vs. number of components, B) scores plot, C) predicted vs. measured plot

RPLC-UV chromatograms (e.g. acetic acid, formic acid, lactic acid, etc.). Anion exchange was employed to retain the dissociated form of these weak acids and elution was achieved by increasing the hydroxide ion concentration in the mobile phase. Figure 3.12 illustrates the IC chromatograms of all samples. Three prominent peaks were observed at 4.9, 5.2, and 6.7 minutes, corresponding to acetic acid, propionic acid, and chloride ion, respectively. Identification of compounds was achieved by comparison with a reference standard. A chemometric model was subsequently developed to correlate IC chromatograms of the 21 hydrolysate samples with inhibition. Sample 20 was identified as an outlier by the model, possibly due to its significantly higher chloride ion concentration compared to other samples. Omission of this sample from the calibration sample set resulted in an improvement in R^2 and RMSEP from 0.50 and 12 to 0.60 and 8, respectively. The regression overview for the developed model is presented in supplementary material presented in Figure 3.13. Overall, the IC model demonstrated a weaker correlation, using two PCs, compared to model 1. The model was subsequently used to predict inhibition of five validation samples and the results are presented in Table 3.2 (column 5). The model had an R^2 of 0.60 and RMSEP of 8. Validation samples 5, 9, and 19 had predicted values that were within 95% confidence intervals of experimentally determined inhibition. Results of the validation step indicate a weaker correlation between ion chromatography data and microbial inhibition, compared to the model constructed with RPLC-UV data. Nevertheless, the regression coefficient plot and the x-loadings plot identified all three retention times as significant variables. The x-loadings plot is depicted in Figure 3.14.

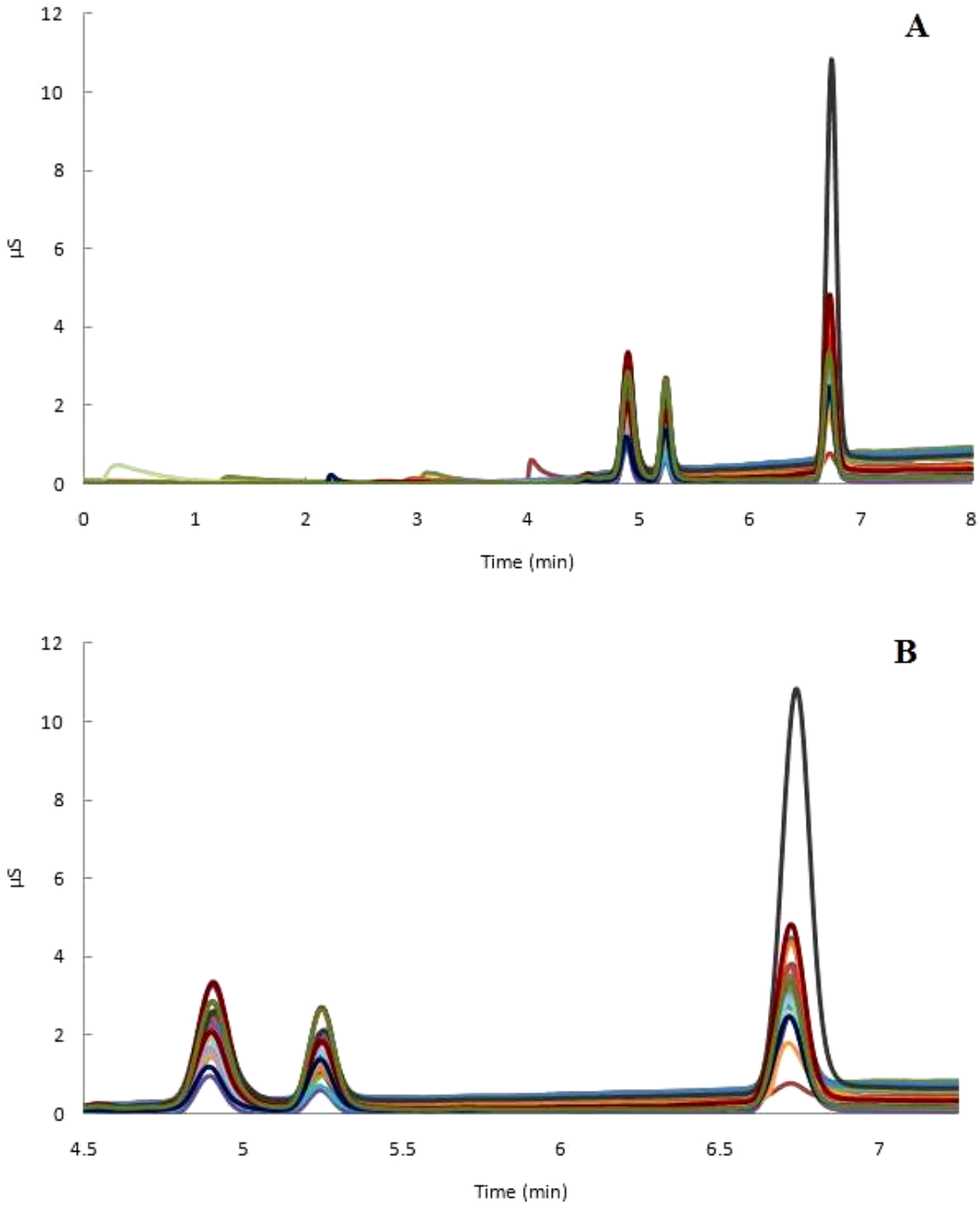


Figure 3.12. IC chromatograms of hydrolysate samples: A) Chromatograms of all hydrolysate samples. B) Expansion of chromatograms shown in A.

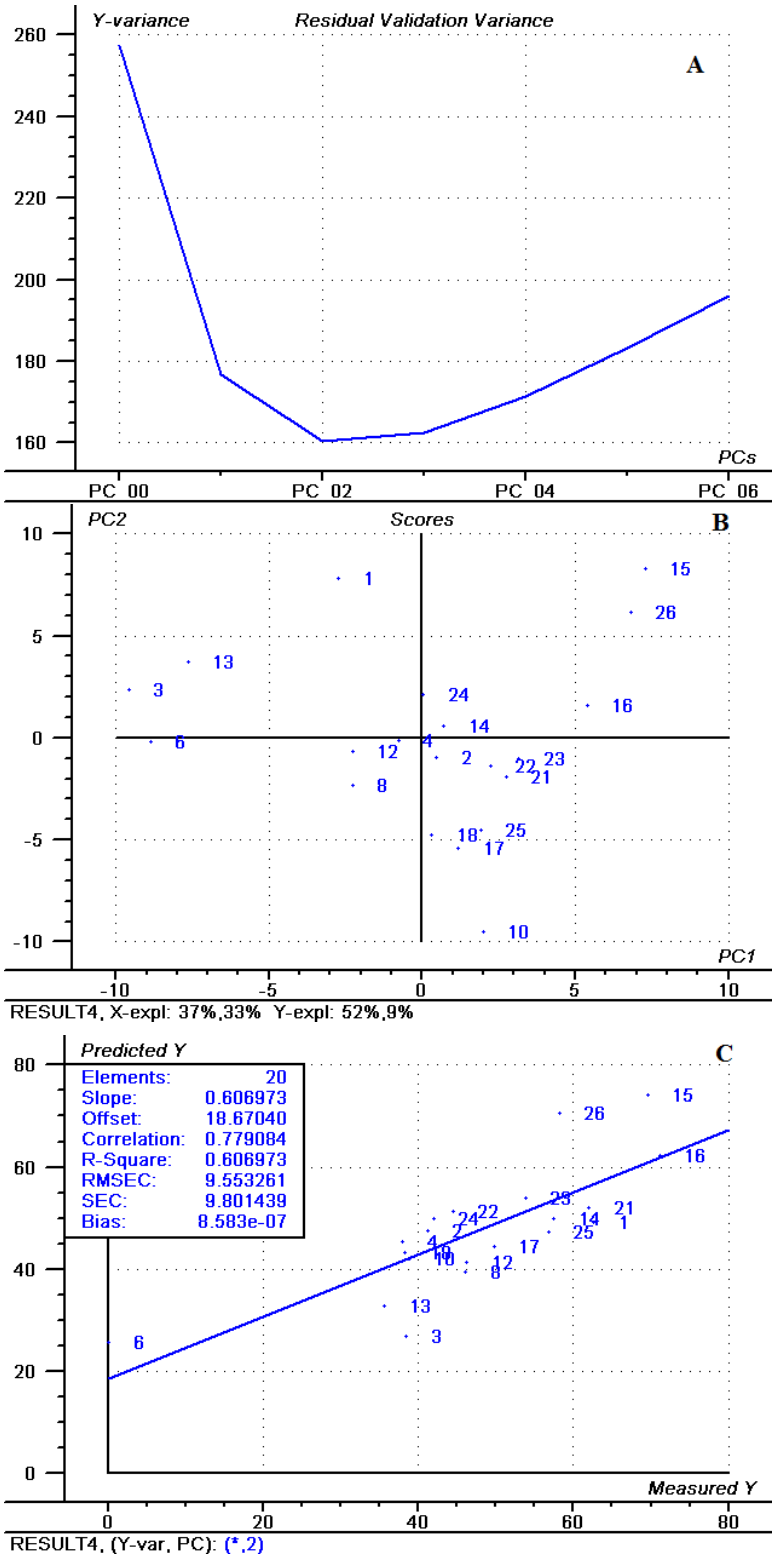


Figure 3.13. Regression overview for model constructed using ion chromatography data: A) variance vs. number of components, B) scores plot, C) predicted vs. measured plot.

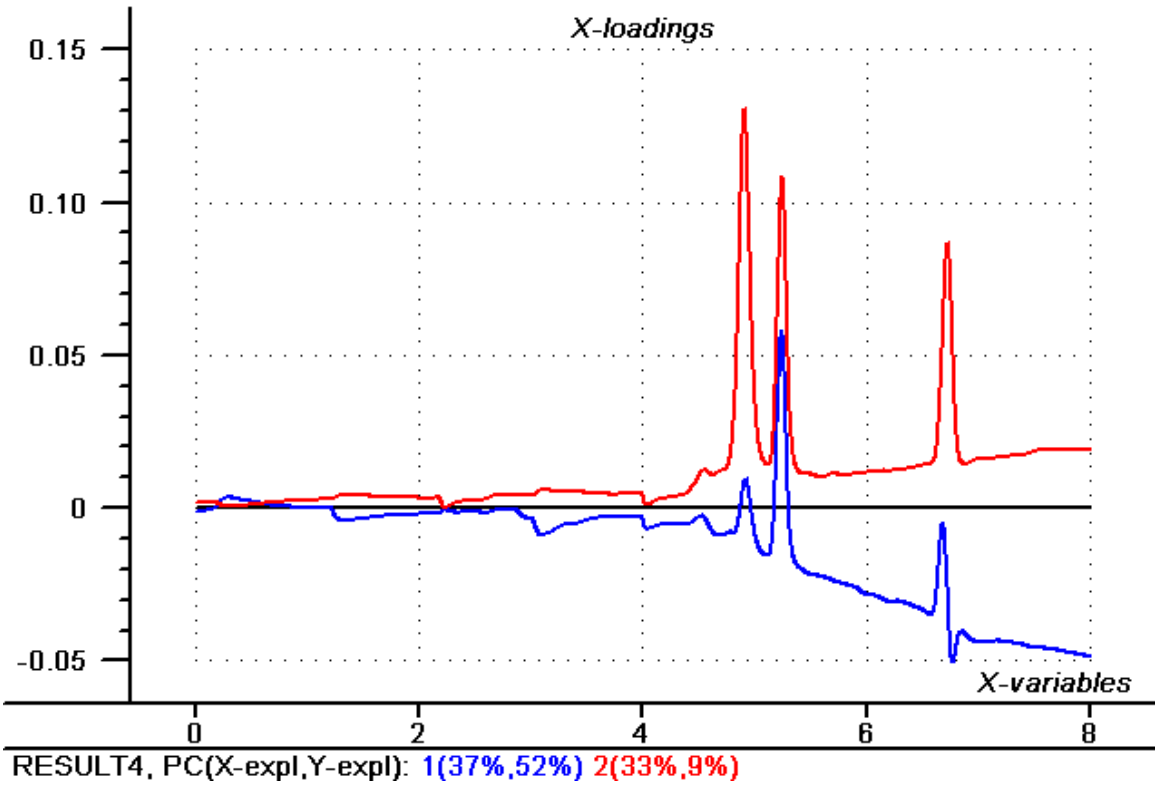


Figure 3.14. X-loadings for ion chromatography model.

Retention time at 4.9 minutes, corresponding to the expected retention time of acetic acid, had the highest loading value, suggesting a stronger correlation with observed inhibition compared to other resolved peaks. Previous studies have demonstrated that acetic acid is inhibitory to ethanol production and cell growth at concentrations of the undissociated form exceeding 5 g/L .[45,46] This is due to the fact that the undissociated acid can diffuse through the plasma membrane and dissociate inside the cell. As a result, protons must be pumped across the membrane to maintain intracellular pH, a process that requires ATP. Above a critical concentration of undissociated acid, the diffusion rate of the undissociated from will exceed the proton pumping capacity of the plasma membrane, leading to cell death. The maximum concentration of acetic acid in hydrolysate samples used in this work was 3 g/L, a value lower than the reported

threshold. However, it should be noted that previous studies were carried out by observing the effects of acetic acid on ethanol production and yeast growth when added alone, and did not investigate its toxicity when present in a mixture of hydrolysate constituents. Therefore, the observed difference between reported inhibitory concentrations of acetic acid and inhibitory concentrations found in hydrolysate samples used in this work may be due to synergistic effects between acetic acid and other degradation compounds. The IC model also identified retention times corresponding to propionic acid and chloride. Chloride has been shown to have detrimental effects when present in concentrations exceeding 6 g/L.[48] The concentration of chloride in hydrolysate samples used in this work ranged from 0.5 to 3.5 g/L. Toxicity studies on propionic acid have not been reported previously. However, other low molecular weight aliphatic acids have demonstrated inhibitory effects on ethanol production.[45-47,112]

It is also important to recognize that although the chemometric model selected the peak corresponding to chloride as a significant variable, it may be the counter cation that has a high correlation with inhibition. Based on previous studies, it is known that potassium ion is present in concentrations that are 10-60 times higher than other cations (Na^+, Ca^{2+}, Mg^{2+}, NH_4^+) in water extracts of corn stover.[20] Thus, it is likely that the counter cation for the identified chloride peak is potassium, which has been shown to inhibit hexokinase, enolase, and membrane ATPase activities and nucleic acid uptake in yeast cells.[49-51] Overall, the results of the above experiment confirm that improved resolution and/or detectability of sample constituents may lead to identification of more retention times as significant variables.

Conclusions

We have successfully demonstrated that chemometric modeling can be used to correlate microbial inhibition with chromatographic data. Furthermore, we have demonstrated that retention times identified by RPLC-UV and IC models as being strongly correlated with inhibition correspond to inhibitory compounds. The utility of this approach may be further improved by better resolution and more universal detectability of sample constituents. Future work should additionally aim to reduce the developed model to practice by expanding it to include more samples from different feedstock types, pretreatment methods and/or fermentative organisms. It is important to note that high correlation does not always suggest causation. As such, any compounds identified as significant by a chemometric model, must be further investigated to confirm their inhibitory effects on microbial processes. This is still most readily accomplished by traditional, bottom-up toxicity testing. However, chemometric modeling of chromatographic data provides a facile approach for identifying candidates and should significantly reduce the number of hydrolysate constituents that need to be tested in order to elucidate the most inhibitory compounds.

CHAPTER FOUR

Conclusions and Final Remarks

We have demonstrated that chemometric modeling coupled with UV-visible and LC-UV analyses can be a powerful tool for prediction of hydrolysate fermentability and identification of fermentation inhibitors arising from pretreatment of biomass. This is a novel method that offers significant increase in throughput for assessment of hydrolysate fermentability and identification of inhibitors when compared to traditional methods. Information obtained from using this methodology can be further used to facilitate the development of inhibitor-resistant microorganisms and/or appropriate detoxification methods. Future work should aim to reduce these models to practice by expanding their sample size to include a wider range of feestock types, pretreatment chemistries, and fermentative microorganisms. Ideally, multiple feedstock types (switchgrass, corn stover, sorghum, poplar, sugarcane bagasse, etc.) should be selected to construct feedstock-specific models. Each feedstock should then be pretreated using various commonly used pretreatment methods (acid hydrolysis, ammonia fiber expansion (AFEX), steam explosion, alkaline treatment, etc.) to account for varying effects of pretreatment chemistry/severity on overall hydrolysate toxicity. Given that the inhibitory effects of degradation compounds will vary for different microorganisms, each hydrolysate should be fermented using various commonly used microorganisms (*E. coli*, *S. cerevisiae*, *Z. mobilis*, *P. stipitis*). Therefore, three different types of models would be constructed: feedstock-specific, pretreatment-specific, and feedstock-pretreatment-

specific models. Ideally, 200-300 total hydrolysate samples should be generated for each feedstock type.

The larger the number of samples that are included in the models, the more representative the models will be of future samples. This will in turn make better prediction models. A schematic of the proposed methodology for two feedstocks is depicted in Figures 4.1-4.5.

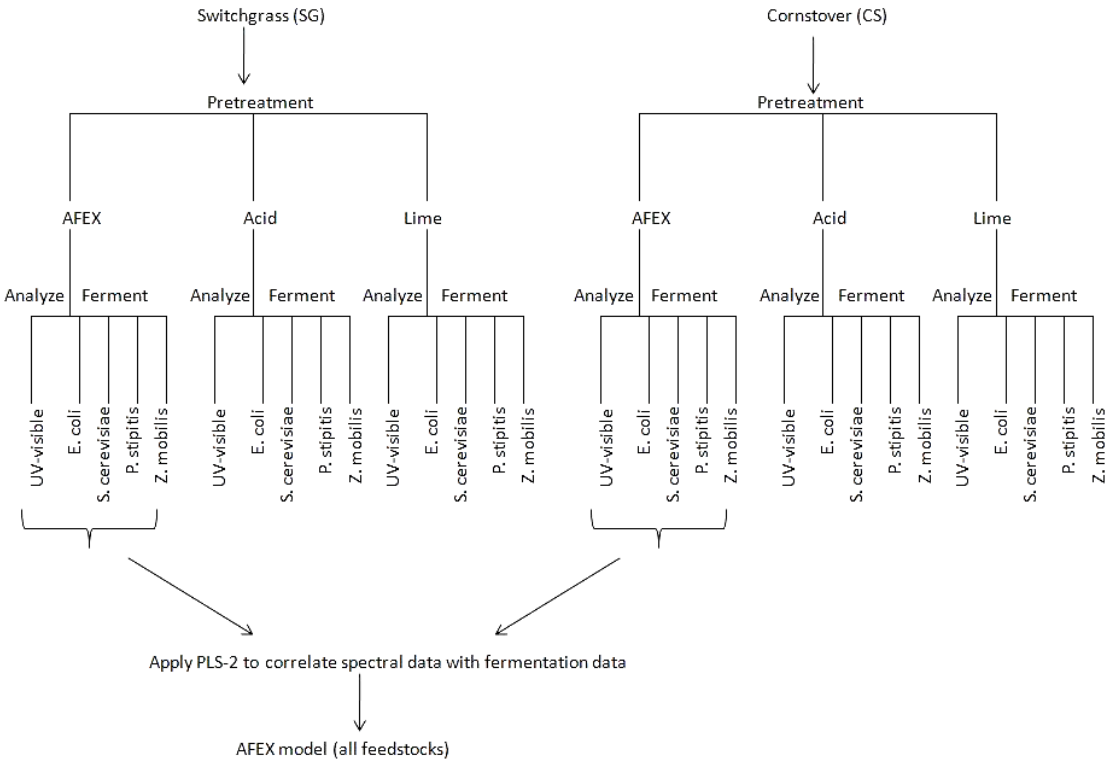


Figure 4.1 Methodology for development of chemometric model utilizing data from all AFEX pretreated hydrolysates.

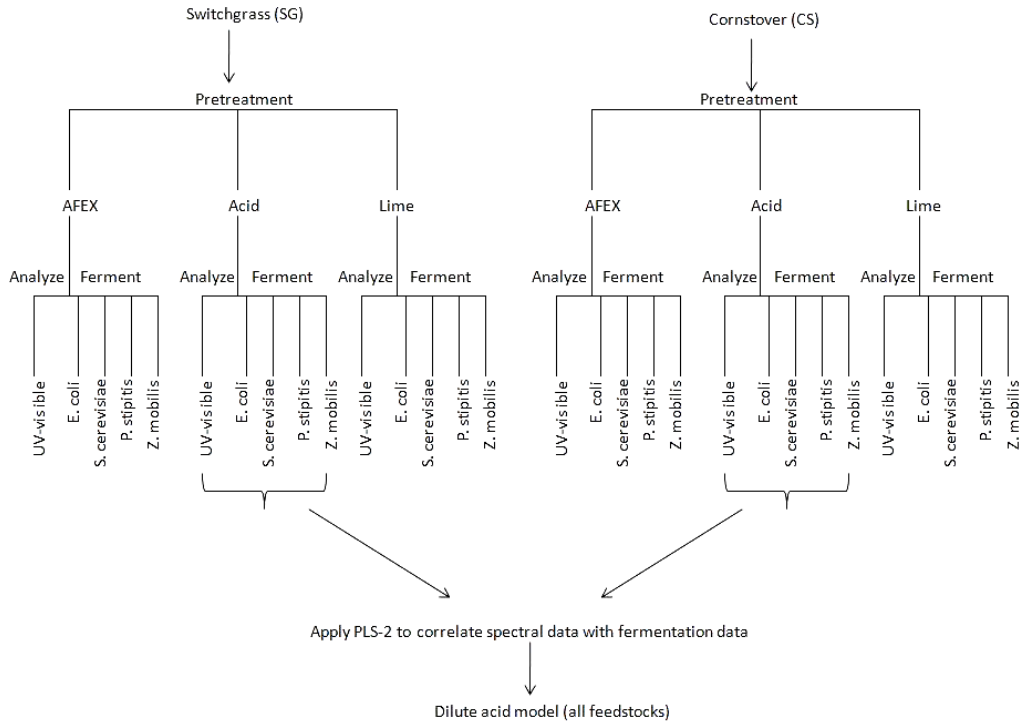


Figure 4.2. Methodology for development of chemometric model utilizing data from all dilute-acid pretreated hydrolysates.

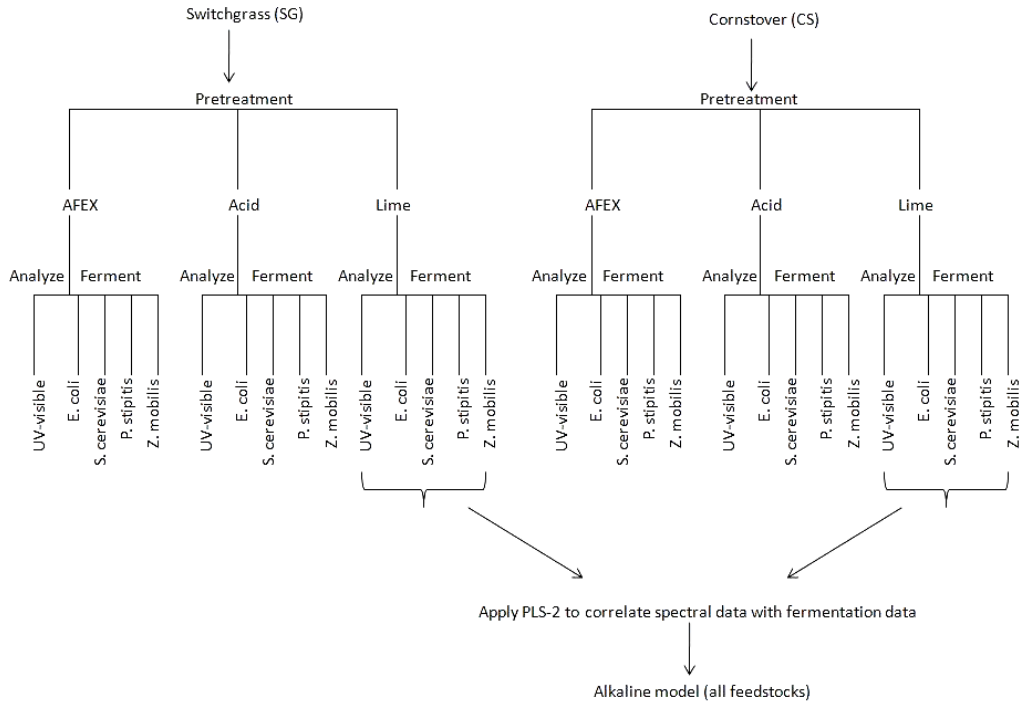


Figure 4.3. Methodology for development of chemometric model utilizing data from all lime pretreated hydrolysates.

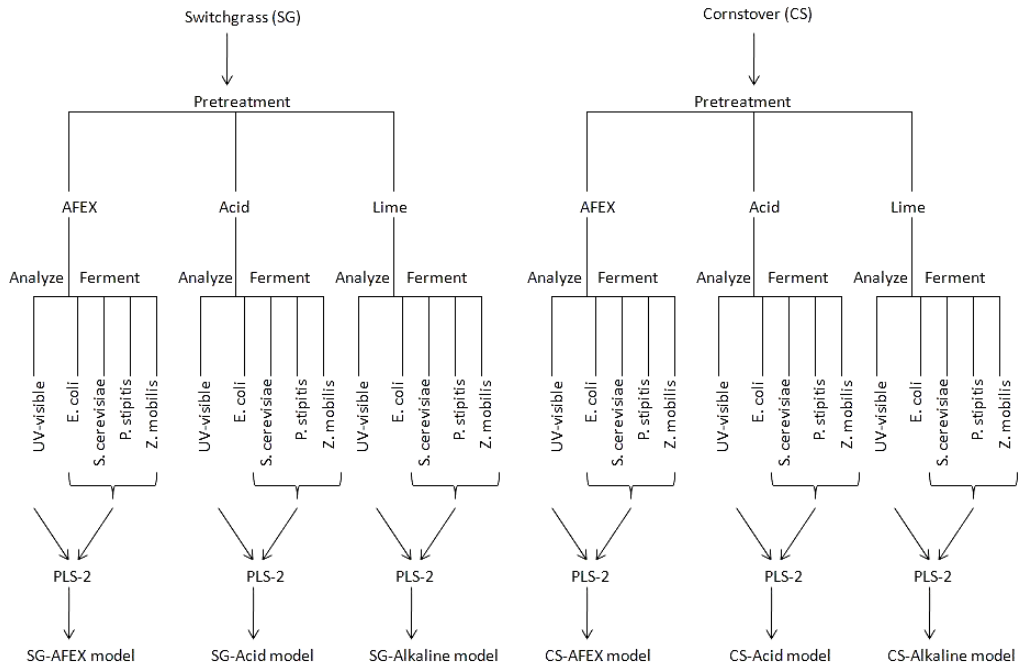


Figure 4.4. Methodology for development of feedstock-pretreatment-specific models.

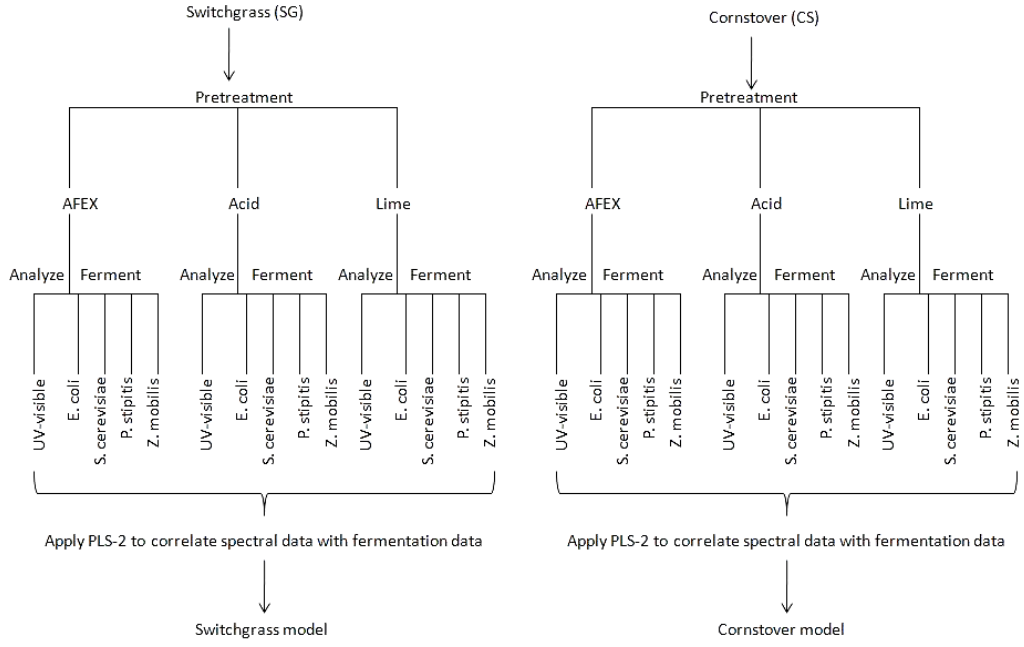


Figure 4.5. Methodology for development of feedstock-specific models.

As previously demonstrated by Morita and coworkers,[101] using a model constructed with data from a specific yeast strain to carry out predictions for a different strain will lead to higher prediction error. Similarly, it would be unlikely that a chemometric model constructed with data from a specific feedstock (e.g. switchgrass) or pretreatment chemistry (e.g. AFEX) could be successfully applied to data obtained from a different set of bioprocessing parameters. Accordingly, one model should be constructed with data from all feedstocks and pretreatment chemistries ("all-in-one" model) (Figure 4.6) and prediction results should be compared with other models in order to select the best prediction model(s).

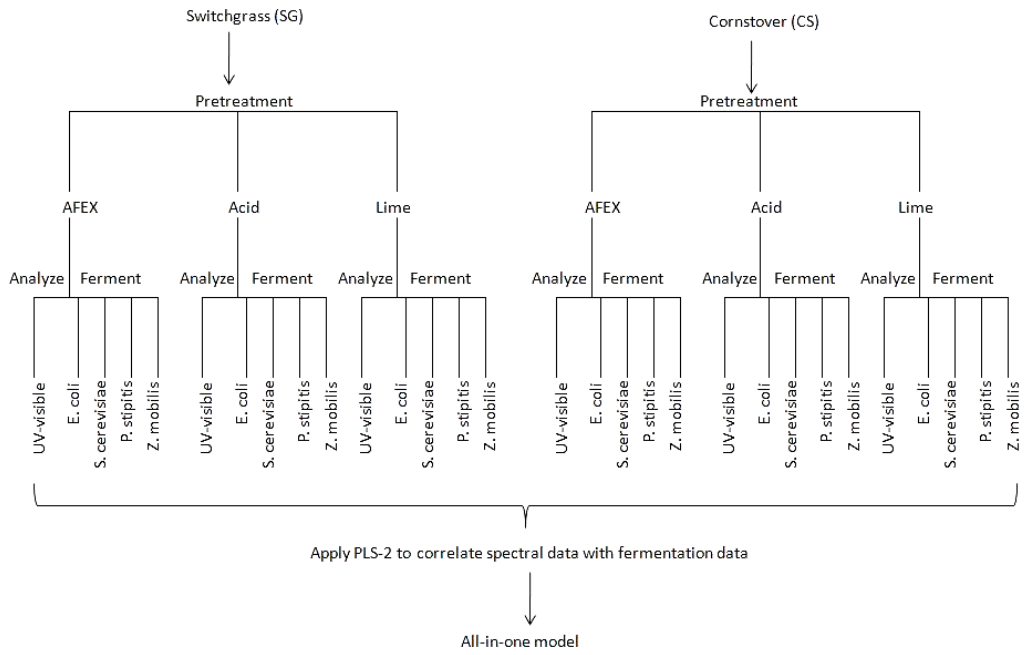


Figure 4.6. Methodology for development of a model with data from all hydrolysates (all-in-one model).

Each model should be validated with samples that are compositionally different from ones used to construct the models, but from the same feedstock type and/or pretreatment

chemistry. Calibration sample size for each model can be increased over time by adding new samples and recalibrating the model.

The HPLC-UV models can be constructed in the same manner as those depicted in Figures 4.1-4.6, with the exception of HPLC-UV being the method of choice for qualitative analysis of hydrolysate. If chromatographic data are obtained in a three-dimensional format (i.e. at multiple wavelengths), Tri-PLS-2 should be used instead of PLS-2. These models can be subsequently used for identification of significant retention times (i.e. compounds) whose change in peak area (i.e. concentration) has the strongest correlation with inhibition. The identified retention times will likely correspond to inhibitory degradation compounds. However, as mentioned before, high correlation does not always suggest causation. Accordingly, all compounds identified as significant by a model will need to be tested to confirm their inhibitory effects on fermentation. The proposed methodology can also be carried using a different detection method to identify additional retention times.

REFERENCES

(1) http://eia.gov/steo/ (accessed July 25, 2011)

(2) Greene, D.; Hopson, J.; Li, J. *Energy and Environmental Concerns* **2004**, 1-9.

(3) Hallock, J. L.; Tharakan, P. J.; Hall, C. A. S.; Jefferson, M.; Wu, W. *Energy* **2009**, *29*, 1673-1696.

(4) Demirbas, A. *Energy Sources* **2005**, *27*, 327-337.

(5) Sun, Y.; Cheng, J. Y. *Bioresour. Technol.* **2002**, *83*, 1-11.

(6) Chum, H. L.; Overend, R. P. *Fuel Process. Technol.* **2001**, *71*, 187-195.

(7) Wu, M.; Wu, Y.; Wang, M. *Biotechnol. Prog.* **2006**, *22*, 1012-1024.

(8) Chandel, A. K.; Chan, E. S.; Rudravaram, R.; Narasu, L. M.; Rao, V. L.; Ravindra, P. *Biotechnology and Molecular Biology Rev.* **2007**, *21*, 14-32.

(9) Shao, X.; Lynd, L.; Bakker, A.; LaRoche, R.; Wyman, C. *Bioprocess Biosyst. Eng.* **2010**, *33*, 485-493.

(10) Wyman, C. E. *Biotechnol. Prog.* **2003**, *19*, 254-262.

(11) van Maris, A. J. A.; Abbott, D. A.; Bellissimi, E.; van den Brink, J.; Kuyper, M.; Luttik, M. A. H.; Wisselink, H. W.; Scheffers, W. A.; van Dijken, J. P.; Pronk, J. T. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology* **2006**, *90*, 391-418.

(12) Sun, Y.; Cheng, J. *Bioresource Technol* **2002**, *83*, 1–11.

(13) Klinke, H. B.; Thomsen, A. B.; Ahring, B. K. *Appl. Microbiol. Biotechnol.* **2004**, *66*, 10–26.

(14) Palmqvist, E.; Hahn-Hägerdal, B. *Bioresource. Technol.* **2000**, *74*, 25-33.

(15) Larsson, S.; Palmqvist, E.; Hahn-Hagerdal, B.; Tengborg, C.; Stenberg, K.; Zacchi, G.; Nilvebrant, N. O. *Enzyme and Microb. Technol.* **1999,** *24*, 151-159.

(16) Du, B.; Sharma, L. N.; Becker, C.; Chen, S-F.; Mowery, R. A.; van Walsum, G. P.; Chambliss, C. K. *Biotechnol. Bioeng.* **2010**, *107*, 430-440.

(17) Sharma, L. N. Identification and Quantitation of Potential Fermentation Inhibitors in Biomass Pretreatment Hydrolysates Using High Performance Liquid Chromatography in Combination with Ultraviolet Detection and Tandem Mass Spectrometry. Ph.D. Dissertation, Baylor University, Waco, TX, 2009.

(18) Hames, B. R.; Thomas, S. R.; Sluiter, A. D.; Roth, C. J.; Templeton, D. W. *Appl. Biochem. Biotechnol.* **2003**, *105-108*, 5-16.

(19) Pordesimo, L. O.; Hames, B. R.; Sokhansanj, S.; Edens, W. C. *Biomass and Bioenergy* **2005**, *28*, 366-374.

(20) Chen, S-F; Mowery, R. A.; Scarlata, C. J.; Chambliss, C. K. *J. Agric. Food Chem.* **2007**, *55*, 5912-5918.

(21) Chen, S-F.; Mowery, R. A.; Sevcik, R. S.; Scarlata, C. J.; Chambliss, C. K. *J. Agric. Food Chem.* **2010**, *58*, 3251-3258.

(22) Sluiter, J. B.; Ruiz, R. O.; Scarlata, C. J.; Sluiter, A. D.; Templeton, D. W. *J. Agric. Food Chem.* **2010**, *58*, 9043-9053.

(23) Tamaki, Y.; Mazza, G. *J. Agric. Food Chem.* **2011**, *59*, 504-512.

(24) DeMartini, J. D.; Studer, M. H.; Wyman, C. E. *Biotechnol. Bioeng.* **2011**, *108*, 306-312.

(25) http://www.nrel.gov/biomass/pdfs/42618.pdf (accessed August, 19, 2011).

(26) Pienkos, P. T.; Zhang, M. *Cellulose* **2009**, *16*, 743-762.

(27) Klinke, H. B.; Ahring, B. K.; Schmidt, A. S.; Thomsen, A. B. *Bioresour. Technol.* **2002**, *82*, 15-26.

(28) Luo, C. D.; Brink, D. L.; Blanch, H. W. *Biomass & Bioenergy* **2002,** *22*, 125-138.

(29) Fenske, J. J.; Griffin, D. A.; Penner, M. H. *J. Ind. Microbiol. Biotechnol.* **1998**, *20*, 364-368.

(30) Chen, S-F; Mowery, R. A.; Castleberry, V. A.; van Walsum, G. P.; Chambliss, C. K. *J. Chromatogr. A* **2006**, *1104*, 54-61.

(31) Sharma, L. N.; Becker, C.; Chambliss, C. K. *Methods in Molecular Biology* **2009**, *581*, 125-143.

(32) Scarlata, C. J.; Hyman, D. A. *J. Chromatogr., A* **2010**, *1217*, 2082-2087.

(33) Boyer, L. J.; Vega, J. L.; Klasson, K. T.; Clausen, E. C.; Gaddy, J. L. *Biomass and Bioenergy*. **1992**, *3*, 41-48.

(34) Palmqvist, E.; Grage, H.; Meinander, N. Q.; Hahn-Hagerdal, B. *Biotechnol. Bioeng.* **1999**, *63*, 46-55.

(35) Zheng-yun, W.; Yu, D.; Li, T.; Yue-hong, L.; Yi-jie, Z.; Wen-xue, Z. *African Journal of Biotechnology* **2010**, *9*, 8661-8666.

(36) Palmqvist, E.; Hahn-Hagerdal, B. *Bioresour. Technol.* **2000**, *74*, 25-33.

(37) Palmqvist, E.; Almeida, J. S.; Hahn-Hagerdal, B. *Biotechnol. Bioeng.* **1999**, *62*, 447-454.

(38) Zaldivar, J.; Martinez, A.; Ingram, L. O. *Biotechnol. Bioeng.* **2000**, *68*, 525-530.

(39) Weigert, B.; Klein, K.; Rizzi, M.; Lauterbach, C.; Dellweg, H. *Biotechnol. Lett.* **1988**, *10*, 895-900.

(40) Gao, J.; Zhang, Y.; Ntoni, J.; Begonia, M. F. T.; Lee, K. S.; Hicks, L.; Hwang, W. W.; Hwang, H-M. *Journal of the Mississippi Academy of Sciences* **2006**, *51*, 220-230.

(41) Modig, T.; Liden, G.; Taherzadeh, M. J. *Biochemical Journal* **2002**, *363*, 769-776.

(42) Martinez, A.; Rodriguez, M. E.; York, S. W.; Preston, J. F.; Ingram, L. O. *Biotechnology Progress* **2000,** *16*, 637-641.

(43) Francisco, R.; Silvia, B.; Gonzalez-Benito, G.; Garcia-Cubero, M. T. *New Biotechnology* **2009**, *25S*, S261.

(44) Zaldivar, J.; Martinez, A.; Ingram, L. O. *Biotechnol. Bioeng.* **1999**, *65*, 24-33.

(45) Taherzadeh, M. J.; Niklasson, C.; Liden, G. *Chemical Engineering Science* **1997,** *52*, 2653-2659.

(46) Casey, E.; Sedlak, M.; Ho, N. W. Y.; Mosier, N. S. *FEMS Yeast Res.* **2010**, *10*, 385-393.

(47) Graves, T.; Narendranath, N. V.; Dawson, K.; Power, N. *J. Ind. Microbiol. Biotechnol.* **2006**, *33*, 469-474.

(48) Doelle M.; Greenfield P. F.; Doelle, H. W. *Process Biochem Int*. **1990**, 151–156.

(49) Hewitt, E. J.; Nicholas, D. J. D. *Metabolic Inhibitors*; Academic Press: New York, 1963.

(50) Kotyk, A.; Rihova, L.; Ponec, M. *Folia Microbiologica* **1971**, *16*, 451-453.

(51) Horak, J.; Rihova, L.; Kotyk, A. *Biochimica et Biophysica Acta* **1981**, *649*, 436-440.

(52) Maiorella, B. L.; Blanch, H. W.; Wilke, C. R. *Biotechnol. Bioeng.* **1984**, *26*, 1155-1166.

(53) Akrida-Demertzi, K.; Demertzis, P. G.; Koutinas, A. A. *Biotechnol. Bioeng.* **1988**, *31*, 666-669.

(54) Casey, E.; Mosier, N. S.; Stockdale, Z.; Ho, N.; Adamec, J.; Sedlak, M. Unpublished work, Purdue University, 2011.

(55) Liu, Z. L.; Slinginger, P. J.; Dien, B. S.; Berhow, M. A.; Kurtzman, C. P.; Gorsich, S. W. *J. Ind. Microbiol. Biotechnol.* **2004**, *31*, 345-352.

(56) Graves, T.; Narendranath, N. V.; Dawson, K.; Power, R. *Appl. Microbiol. Biotechnol.* **2007**, *73*, 1190-1196.

(57) Helle, S. S.; Cameron, D. R.; White, B.; Lam, J.; Sheldon, J. B. *Enz. Microb. Technol.* **2003**, *33*, 786-792.

(58) Oliva, J. M.; Ballesteros, I.; Negro, M. J.; Manzanares, P.; Cabanas, A.; Ballesteros, M. *Biotechnol. Prog.* **2004**, *20*, 715-720.

(59) Oliva, J. M.; Negro, M. J.; Saez, F.; Ballesteros, I.; Manzanares, P.; Gonzales, A.; Ballesteros, M. *Process Biochem.* **2006**, *41*, 1223-1228.

(60) Cao, G-L; Ren, N-Q; Wang, A-J; Guo, W-Q; Xu, J-F; Liu, B-F. *Int. J. Hydrogen Energy* **2010**, *35*, 13475-13480.

(61) Klinke, H. B.; Thomsen, A. B.; Ahring, B. K. *Appl. Microbiol. Biotechnol.* **2001**, *57*, 631-638.

(62) Olsson, L.; HahnHagerdal, B. *Enz. and Microb. Technol*. **1996**, *18*, 312-331.

(63) Wang, B.; Feng, H. 2010. Detoxification of Lignocellulosic Hydrolysates. In *Biofuels from Agricultural Wastes and Byproducts*; Blaschek, H. P., Ezeji, T. C, Scheffran, J. Eds. Blackwell Publishing: New Jersey, 2010.

(64) Liu, Z. L. *Appl. Microbiol. Biotechnol.* **2011**, *90*, 809-825.

(65) Jonsson, L. J.; Palmqvist, E.; Nilvebrant, N. O.; Hahn-Hagerdal, B. *Microbiol. Biotechnol.* **1998**, *49*, 691-697.

(66) Lopez, M. J.; Nichols, N. N.; Dien, B. S.; Morena, J.; Bothast, R. J. *Appl. Microbiol. Biotechnol.* **2004**, *64*, 125-131.

(67) Nichols, N. N.; Dien, B. S.; Guisado, G. M.; Lopez, M. J. *Appl. Biochem. Biotechnol.* **2005**, *121-124*, 379-390.

(68) Nichols, N. N.; Sharma, L. N.; Mowery, R. A.; Chambliss, C. K.; van Walsum, G. P.; Dien, B. S.; Iten, L. B. *Enz. Microb. Technol.* **2008**, *42*, 624-630.

(69) Parajo, J. C.; Dominguez, H.; Dominguez, J. M. *Bioprocess Eng.* **1995**, *16*, 39-43.

(70) Zyl, C. V.; Prior, B. A.; Preez, J. C. D. *Enz. Microb. Technol.* **1991**, *13*, 82-86.

(71) Agblevor, F. A.; Fu, J.; Hames, B.; McMillan, J. D. *Appl. Biochem. Biotechnol.* **2004**, *119*, 97-120.

(72) Esbensen, K. H. In *Multivariate Data Analysis-in Practice*, 5th ed.; Guyot, D., Westad, F., HoumÃƒÂ¸ller, L. P., Eds.; Camo: New Jersey, 2010.

(73) Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics: A Practical Guide*; John Wiley & Sons: New York, 1998.

(74) Wold, S. *Chemometrics and Intelligent Laboratory Systems* **1995,** *30*, 109-115.

(75) Davis, C. B. Application of Chemometric Analysis to UV-Visible and Diffuse Near-Infrared Reflectance Spectra. Ph.D. Dissertation, Baylor University, Waco, Texas, 2007.

(76) Davis, C. B.; Markey, C. E.; Busch, M. A.; Busch, K. W. *J. Agric. Food Chem.* **2007**, *55*, 5925-5933.

(77) Otto, M. *Chemometrics: Statistics and Computer Application in Analytical Chemistry*; Wiley-VCH: Germany, 2007.

(78) Wold, S.; Sjostrom, M. *ACS Symp. Ser.* **1977**, *52*, 243-282.

(79) Dogra, J. A. Multivariate Analyses of Near –Infrared and UV Spectral Data. Ph.D. Dissertation, Baylor University, Waco, Texas, 2009.

(80) Hsiao, C. P.; Siebert, K. J. *Internation Journal of Food Microbiology* **1999**, *47*, 189-201.

(81) de la Mata-Espinosa, P.; Bosque-Sendra, J. M.; Bro, R.; Cuadros-Rodriguez, L. *Anal. Bioanal. Chem.***2011**, *339*, 2083-2092.

(82) Dunlop, P. J.; Bignell, C. M.; Jackson, J. F.; Hibbert, D. B.; *Chemom. Intell. Lab. Syst.* **1995**, *30*, 59-67.

(83) Schievano, E.; Peggion, E.; Mammi, S. **2010**, *58*, 57-65.

(84) Lonni, A. A. S. G.; Scarminio, I. S.; Silva, L. M. C.; Ferreira, D. T. *Analytical Sciences* **2003**, *19*, 1013-1017.

(85) Galtier, O.; Dupuy, N.; Le Dreau, Y.; Ollivier, D.; Pinatel, C.; Kister, J.; Artaud, J. *Anal. Chim. Acta* **2007**, *595*, 136-144.

(86) Donmez, O. A.; Bozdogan, B. A. A.; Sungur, S. *Talanta* **2011**, *83*, 1601-1605.

(87) Sikorska, E.; Gliszczyńska-Świgło, A.; Insińska-Rak, M.; Khmelinskii, I.; De Keukeleire, D.; Sikorski, M. *Anal. Chim. Acta* **2008**, *613*, 207-217.

(88) Brereton, R. G. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant;* John Wiley & Sons: West Sussex, 2003; pp 221-324.

(89) Tran, A. V.; Chambers, R. P., *Enz. Microb. Technol.* **1986**, *8*, 439-444.

(90) Clark, T. A.; Mackie, K. L., *Journal of Chemical Technology and Biotechnology B-Biotechnology* **1984**, *34*, 101-110.

(91) Pfeifer, P. A.; Bonn, G.; Bobleter, O., *Biotechnology Letters* **1984**, *6*, 541-546.

(92) Taherzadeh, M. J.; Eklund, R.; Gustafsson, L.; Niklasson, C.; Liden, G. *Industrial & Engineering Chemistry Research* **1997**, *36*, 4659-4665.

(93) http://energy.senate.gov/public/index.cfm?FuseAction=IssueItems.Detail&IssueItem_ID=f10ca3dd-fabd-4900-aa9d-c19de47df2da&Month=12&Year=2007 (accessed 08/16/2011).

(94) Lynd, L. R.; *Annual Review of Energy and the Environment* **1996**, *21*, 403-465.

(95) Galbe, M.; Zacchi, G. *Appl. Microbiol. Biotechnol*. **2002**, *59*, 618-628.

(96) Klinke, H. B.; Olsson, L.; Thomsen, A. B.; Ahring, B. K. *Biotechnol. Bioeng.* **2003**, *81*, 738-747.

(97) McMillan, J. D. In *Enzymatic Conversion of Biomass for Fuels Production*; Himmel, M. E., Baker, J. O., Overend, R. P. Eds.; American Chemical Society: Ohio, 1994.

(98) Persson, P.; Andersson, J.; Gorton, L.; Larsson, S.; Nilvebrant, N-O; Jonsson, L. J. *J. Agric. Food Chem.* **2002**, *50*, 5318-5325.

(99) Overend, R. P.; Chornet, E. *Philosophical Transactions of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences.* **1987**, *321*, 523-536.

(100) Chen, S-F.; Mowery, R. A.; Chambliss, C. K.; van Walsum, P. G. *Biotechnol. Bioeng.* **2007**, *98*, 1135-1145.

(101) Morita, H.; Hasunuma, T.; Vassileva, M.; Tsenkova, R.; Kondo, A. *Analytical Chemistry*. **2011**, *83*, 4023-4029.

(102) http://www.nrel.gov/biomass/pdfs/42630.pdf (accessed June 16, 2010).

(103) Sevcik, R. S.; Mowery, R. A.; Becker, C.; Chambliss, C. K. *J. Chromatogr., A* **2011**, *1218*, 1236-1243.

(104) Holler, F. J.; Crouch, S. R.; Skoog, D. A. *Principals of Instrumental Analysis*; Cengage Learning: Kentucky, 2007.

(105) Nakai, S. A.; Siebert, K. J. *International Journal of Food Microbiology* **2003**, *86*, 249-255.

(106) Lorber, A.; Wangen, L. E.; Kowalski, B. R. *Journal of Chemometrics* **2005**, *1*, 19-31.

(107) Kucukboyaci, N.; Aysegul, G.; Erdal, D.; Adiguzel, N.; Bani, B. *Journal of Separation Science* **2010**, *33*, 2558-2667.

(108) Maggio, R. M.; Damiani, P. C.; Olivieri, A. C. *Talanta* **2011**, *83*, 1173-1180.

(109) Bas, D.; Boyaci, I. H. *Journal of Food Engineering* **2007**, *78*, 836-845.

(110) Sivakumar, T.; Manavalan, R.; Muralidharan, C.; Valliappan, K. *Journal of Pharmaceutical and Biomedical Analysis* **2007**, *43*, 1842-1848.

(111) Hedayatifar, N.; Rabbe, D. H.; Busch, K. W.; Chambliss, C. K. Baylor University, 2011, unpublished results.

(112) Zaldivar, J.; Ingram, L. O. *Biotechnol. Bioeng.* **1999**, *66*, 203-210.