Multivariate Characterization, Modeling, and Design of Chemical Products in Property Cluster Space

by

Subin Hada

A dissertation submitted to the graduate faculty of
Auburn University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Auburn, Alabama
August 3, 2013

Keywords: Chemical Product Design, Multivariate Statistical Analysis, Chemometric Technique, Reverse Problem Formulation, Property Clustering.

Approved by

Mario R. Eden, Chair and McMillan Professor of Chemical Engineering
Christopher B. Roberts, Uthlaut Professor of Chemical Engineering
Elizabeth Lipke, Assistant Professor of Chemical Engineering
Sushil Adhikari, Assistant Professor of Biosystems Engineering
Mahmoud El-Halwagi, McFerrin Professor of Chemical Engineering

ABSTRACT

The focus of this dissertation is on the development of *in silico* approaches for the logical and systematic solution of chemical product design problems. The application of multivariate characterization, modeling, and design is accomplished by utilizing interdisciplinary methods and tools that extend through multivariate statistics, applied mathematics and computer science. Methodologies and techniques such as spectroscopy-based group contribution methods, chemometric/chemoinformatic techniques, reverse problem formulation, and property clustering techniques are integrated within computer-aided molecular/mixture design (CAMD) algorithms to design chemical products in a computationally efficient manner that provides optimum performance in terms of customer requirements. Property-based design techniques and multivariate data-driven modeling and optimization strategies are presented in this dissertation covering two specific areas of chemical product design: mixture and molecular design.

In mixture design, the property integration framework is combined with multivariate statistical techniques and applied in a reverse problem formulation on chemical product design problems by systematic and insightful use of past data describing the properties of the raw materials, their blend ratios, and the process conditions during the production of a range of product grades to achieve new and improved products. Projection methods, like principal component analysis (PCA) and partial least squares (PLS) are applied to identify the underlying relationships necessary for simultaneous optimization of all three variables. The method is illustrated using a polymer blending problem.

In molecular design, multivariate characterization techniques like infrared (IR) spectroscopy are utilized to generate numerical descriptors of molecular architecture in terms of IR frequency of a set of representative samples. Models based on quantitative structure-property relations (QSPR) are used to elucidate structure-property relationships. Applying principal component analysis, high dimensional and highly correlated molecular descriptor variables are transformed into low dimensional and statistically independent latent variables. These latent variables are then used to calibrate latent property models. Finally, the reverse design of molecules is accomplished by exhaustively searching for molecular structures with target properties, from the combinatorial building blocks. A characterization-based group contribution method is utilized to estimate the properties of the formulated chemical products. The concepts and the solution methodologies are demonstrated using two proof-of-concept examples: biodiesel additive formulation and ionic liquid design.

ACKNOWLEDGEMENTS

My deepest gratitude to my advisor, Dr. Mario R. Eden, for his gentle guidance, patience, and constant encouragement throughout my doctoral program at Auburn University. His mentorship and support has been invaluable on both an academic and a personal level. I will miss travelling with him around the world for conferences.

It is a pleasure to thank my dissertation committee, and outside reader, respectively: Dr. Christopher Roberts, Dr. Elizabeth Lipke, Dr. Sushil Adhikari, Dr. Mahmoud El-Halwagi, and Dr. Steven Taylor. Thank you for having patience as I endeavored to complete my work. I appreciate the feedback and review that helped to improve and complete this dissertation.

I thank my fellow labmates in the Eden laboratory who supported me and shared their enthusiasm for and comments on my work. My special thanks to Robert Herring for the stimulating discussions and for the sleepless nights we worked together before deadlines.

I gratefully acknowledge the financial support provided by the Southeastern Regional Sun Grant Program, and the USDA-AFRI (Award# 2011-67009-20077). In addition, I thank Dr. Koji Muteki; Pfizer Inc., CT, USA, for providing the starch blending data used in the mixture design proof-of-concept example.

I cannot end without thanking my parents, Rajendra P. Hada and Subrata Hada, and my brother Sujin Hada, on whose constant encouragement and love I have relied on throughout long years of my educational journey. Their persistent audacity and conviction will always inspire me. It is to them that I dedicate this work. I would also like to thank my wife, Fui Chi Yap, for her unequivocal support.

TABLE OF CONTENTS

Page

ABSTRACT……ii

ACKNOWLEDGEMENTS ……iv

TABLE OF CONTENTS ……v

LIST OF TABLES……viii

LIST OF FIGURES……x

LIST OF ABBREVATIONS……xiv

CHAPTER 1 INTRODUCTION …… 1

1.1 Chemical Product Design …… 1
1.2 Challenges and motivations …… 4
1.3 Scope and Objectives…… 11
1.4 Significance of the Research…… 12
1.5 Organization…… 13

CHAPTER 2 THEORY AND METHODOLOGY…… 15

2.1 Process-Product Design …… 15
2.2 Property-Based Process and Product Design…… 16
2.2.1 Reverse Problem Formulation …… 17
2.2.2 Property Clustering Technique…… 20
2.3 Prediction of Properties …… 25
2.3.1 Group Contribution Method…… 27
2.4 Characterization Techniques …… 29
2.4.1 Spectroscopy …… 31
2.5 General Regression Models …… 34
2.6 Latent Variable Models …… 36
2.6.1 Principal Component Analysis (PCA)…… 39

2.6.2 Principal Component Regression (PCR) 46
2.6.3 Partial Least Squares (PLS) 47
2.6.4 Model Validation 49
2.7 Principal Properties in Cluster Space 51
2.8 Computer-Aided Design using QSPR and cGCM 52
2.8.1 Solution to Reverse Problem 53
2.8.2 Solution to Forward Problem 54
CHAPTER 3 MIXTURE DESIGN 58
3.1 Introduction 58
3.2 Traditional Approach 59
3.3 Multi-Block Data Structure 62
3.4 Multi-Block Regression Models 63
3.5 Proof of Concept Example – Starch Blending 66
3.5.1 Structure of starch blending data 66
3.5.2 Data analysis 70
3.5.3 Model development 75
3.5.4 Design of desired products in score space 77
3.5.5 Design of desired products in cluster space 79
3.6 Conclusion 83
CHAPTER 4 DESIGN OF BIODIESEL ADDITIVES 85
4.1 Introduction 85
4.2 Structure Property Relationships 89
4.3 Technical Difficulties with Biodiesel Use 90
4.4 Fuel Additives 93
4.5 Additive Design 93
4.5.1 Types of Additives 94
4.5.2 Additive Property Estimation 95
4.5.3 Characterization of the Additive Molecules 97
4.5.4 Additive IR Data Analysis 98
4.5.5 Latent Variable Model Development 100
4.5.6 Translating Physical properties to Latent Properties 102
4.5.7 Evaluation of Desired Additive Feasibility Region 104
4.5.8 Enumeration of Desired Additive Molecules 106
4.6 Conclusion 112

CHAPTER 5 REVERSE DESIGN OF IONIC LIQUIDS 113

5.1 Introduction 113

5.2 Density Functional Theory 115

5.3 Data Analysis and Model Development 120

5.4 Reverse Design of Ionic Liquids using QSPR and cGCM .. 123

5.5 Conclusion 130

CHAPTER 6 FUTURE WORK 132

6.1 Methodology Improvements 132

6.1.1 Multi-Dimensional Characterization 132

6.1.2 Multi-Way Modeling 134

6.1.3 Stochastic Search and Optimization 136

6.1.4 Managing and Handling Uncertainty 137

6.2 Design of Inherently Benign Chemical Process Routes 138

REFERENCES. 141

APPENDIX A MIXTURE MODELS 150

A.1 Scheffe Mixture Model 150

A.1 Cox Mixture Model 151

APPENDIX B SPECTRAL INTERPRETATION 153

B.1 Infrared Spectroscopy 154

B.2 Molecular Vibrational Spectroscopy 155

B.3 Near Infrared Spectroscopy 160

B.4 Characterizing IR Spectroscopy 161

LIST OF TABLES

Table		Page
Table 3.1:	Starch material property data matrix.	67
Table 3.2:	Blending ratio and process condition data matrix.	68
Table 3.3:	Mixture product quality data matrix.	69
Table 3.4:	PCA score values for X- and Y-blocks.	74
Table 3.5:	PCA loading values for X- and Y-blocks.	75
Table 3.6:	The regression coefficients for the expressions in Eqs. (3.13) through Eqs. (3.15).	76
Table 3.7:	Score values for desired products in Figure 3.11.	78
Table 3.8:	Desired product properties for desired products in Figure 3.11.	78
Table 3.9:	Required mixture conditions to achieve target product properties.	79
Table 3.10:	Candidate ternary mixtures and fractional contributions of the constituents.	82
Table 4.1:	Chemical structures of common fatty acids.	87
Table 4.2:	Fatty acid profiles of some common biodiesel feedstock	88
Table 4.3:	Fuel properties of biodiesel fuels and diesel	89
Table 4.4:	Specifications for biodiesel and diesel	92
Table 4.5:	Commercially available diesel additives and their estimated properties.	97
Table 4.6:	Model coefficients using PCR.	101
Table 4.7:	Biodiesel target properties.	102
Table 4.8:	Physical and latent properties describing feasibility region.	103
Table 4.9:	Biodiesel target latent properties.	103
Table 4.10:	Crude biodiesel properties.	103

Table 4.11: Biodiesel additive latent property feasibility region. 105
Table 4.12: Molecular groups and their latent property contributions. 107
Table 4.13: Results from characterization-based molecular design. 109
Table 4.14: Additive solubility in FAME at 25°C. 110
Table 5.1: B3LYP/6-311+G(2d,p) vibrational assignments (cm^{-1}) of $[emIm]PF_6$. 118
Table 5.2: PCA score values for X- and Y-blocks. 121
Table 5.3: Model coefficients using PCR. 122
Table 5.4: Target ionic liquid properties. 124
Table 5.5: Anion groups and their latent property contributions. 125
Table 5.6: Cation groups and their latent property contributions. 126
Table 5.7: Alkyl chain attached to cation groups and their latent property contributions. 127
Table 5.8: Thirteen candidate ionic liquid molecules that match target properties in property space. 130
Table B.1: IR spectrum absorption for different bond types. 160
Table B.2: IR absorbance frequencies and magnitudes of functional groups. 161
Table B.3: Abbreviations, names and structures of investigated ionic liquid training set. 173
Table B.4: Candidate ionic liquid solutions enumerated from exhaustive search in latent property space. 179

LIST OF FIGURES

Figure **Page**

Figure 1.1: Stages involved in chemical product design. 2

Figure 1.2: Chemically formulated products. 3

Figure 1.3: Hierarchy of chemical and biological informatics. 9

Figure 2.1: Conventional solution approach for process and molecular design problems 16

Figure 2.2: Product and process design problem using reverse problem formulation methodology. 18

Figure 2.3: A multi-scale product design framework showing (I) traditional approach and (II) RPF approach linking each of the scales via a common property domain. 19

Figure 2.4: Representation of intra- and inter-stream conservation of clusters in ternary diagram. 22

Figure 2.5: Representation of feasibility reason with source-sink mapping using clusters in ternary diagram. The clustering points are converted from ternary to Cartesian coordinate. 24

Figure 2.6: Classification of property estimation methods. 25

Figure 2.7: QSAR model development steps. 26

Figure 2.8: An overview of the interconnectivity of characterization techniques, molecular architecture, and physical properties and attributes of chemical and material products 30

Figure 2.9: IR spectra of butylated hydroxytoluene molecule 31

Figure 2.10: Unit variance scaled and mean-centered variables. 38

Figure 2.11: Projection of higher dimensional data onto a lower dimensional subspace. 40

Figure 2.12: PCA decomposition of X matrix. 42

Figure 2.13: Scree plot of the correlation matrix. 43

Figure 2.14: Principal component analysis: Score plot (a, left) of t_1/t_2 and loading plot (b, right) of p_1/p_2. The ellipse represents the Hotelling T^2 with 95% confidence in score plot. 45

Figure 2.15: PLS regression on descriptive (X) and response (Y) variables.... 47

Figure 2.16: Forward and reverse problems in computer-aided molecular design. 52

Figure 2.17: An overview on the methodology of multivariate characterization, modeling, and design. 57

Figure 3.1: Traditional approaches in mixture design. 61

Figure 3.2: Data structure for three manipulative variable matrixes and a quality/response variable matrix..... 63

Figure 3.3: Data structure for combined manipulative variable matrixes and a quality/response variable matrix. 64

Figure 3.4: Data structure of X and R matrix for two classes of raw materials and their blend ratios..... 67

Figure 3.5: Distribution plot, outlier box plot and normal quantile plot..... 70

Figure 3.6: Scree plot and pareto plot for PCA on X-variables. 71

Figure 3.7: Combined PCA score and loading plots (Biplot) on first and second components for X-block. 72

Figure 3.8: Combined PCA score and loading plots (Biplot) on first and second components for Y-block. 72

Figure 3.9: PLS model coefficients for blend property matrix Y using four latent factors. 73

Figure 3.10: Predicted vs. actual product properties using PCR model. 77

Figure 3.11: Visualization of target product properties in score space. 78

Figure 3.12: Visualization of starch blending formulation in cluster space... 80

Figure 3.13: Visualization of starch blending formulation in cluster spac 81

Figure 4.1: Overall stoichiometric transesterification reaction scheme..... 86

Figure 4.2: Stearic acid methyl ester. 87

Figure 4.3: A typical triglyceride molecule with different fatty acid chains of soybean oil. 88

Figure 4.4: Oleic acid methyl ester..... 90

Figure 4.5: Compositions of fats and oils and their effects on the fuel properties..... 91

Figure 4.6: Approaches to improving biodiesel fuel properties..... 92

Figure 4.7. Target specifications for bio-diesel and its blend in terms of cetane number, melting point, and kinematic viscosity..... 94

Figure 4.8: Infrared spectra of diesel additive molecules..... 98

Figure 4.9: Scree plot for PCA on additive IR data. 99

Figure 4.10: PCA score plots on first third PCs for additive IR data. 99

Figure 4.11: Predicted vs. actual product properties using PCR model. 101

Figure 4.12: Target feasibility region and crude biodiesel in cluster space. 104

Figure 4.13: Desired additive design feasibility region in cluster space. 105

Figure 4.14: Cluster diagram for biodiesel blending problem..... 108

Figure 4.15: Spatial representation of candidate additive molecules according to B3LYP/6311++G(3df,3dp) calculations: (a) isopropenyl acrylate (CM1), and (b) ethyl acetate (CM3)..... 111

Figure 5.1: Selection of anions, cations, and side chains attached in cations for a task specific ionic liquid application. 114

Figure 5.2: Geometry optimized molecular structure of [emIm]PF_6 at the B3LYP/6-311G(2d,p) computational level..... 116

Figure 5.3: (a) Calculated vs. experimental IR frequencies and (b) infrared spectrum at 2cm^{-1} resolution of [emIm]PF_6 at the B3-LYP/6-311G(2d,p) computational level..... 117

Figure 5.4: Predicted vs. actual IL properties using PCR model..... 122

Figure 5.5: Scenario of reverse design of ILs. 128

Figure 5.6: Enumerated and validated candidate IL molecules. 129

Figure 6.1: Improvements in the process of multivariate characterization, modeling, and design. 133

Figure 6.2: The decomposition of X-block by (a) PCA and (b) PARAFAC. . 135

Figure 6.3: Structure of an artificial neural network..... 136

Figure 6.4: Aspects of early process design..... 138

Figure B.6.5: IR regions of the electromagnetic spectrum. 154

Figure B.6.6: Stretching and bending vibrational modes for H_2O. 155

Figure B.6.7: Stretching and bending vibrational modes for a CH_2 goup. .. 157
Figure B.6.8: Ball and spring model for atoms and bonds respectively. 158
Figure B.6.9: Vibrational bands in infrared spectrum. 160

LIST OF ABBREVATIONS

ADMET	Absorption, Distribution, Metabolism, elimination, and Toxicity
ANN	Artificial Neural Network
AUP	Augmented Property Index
ASTM	America Society for Testing and Materials
BIC	Bayesian Information Criterion
BHT	Butylated Hydroxytoluene
CAMD	Computer Aided Molecular Design
cCAMD	Characterization-based Computer-Aided Molecular Design
cGCM	Characterization-based Group Contribution Method
CI	Connectivity Indices
CN	Cetane Number
CFPP	Cold Filter Plugging Point
CP	Cloud Point
DFT	Density Functional Theory
EB	Elongation at Break
EEM	Emission-Excitation Matrix
EG	Ethylene Glycol
EGMEA	Ethylene Glycol Methyl Ether Acrylate
DOE	Design of Experiments
DG	Diglyceride
DTBP	Di-Tert-Butyl Peroxide
EHS	Environment, Health, and Safety
FAME	Fatty Acid Methyl Ester

FBN	Free Bond Number
GA	Genetic Algorithm
GC	Group Contribution
GCM	Group Contribution Methods
GL	Glycerol
IBI	Inherent Benign-ness Indicator
ILs	Ionic Liquids
IPA	Iso Propyl Alcohol
IR	Infra-Red
LCA	Life Cycle Assessment
LP	Linear Programs
LVs	Latent Variables
MSA	Mass Separating Agent
MC	Monte Carlo
MD	Molecular Dynamics
MG	Monoglyceride
MINLP	Mixed Integer Non-Linear Programs
MLR	Multi-Linear Regression
MM	Molecular Mechanics
MO	Methyl Oleate
MTBE	Methyl Tert-Butyl Ether
NLP	Non Linear Programs
NMR	Nuclear Magnetic Resonance
NIR	Near Infra-Red
NIST	National Institute of Standard and Technology
NPCT	Nonlinear Principal Component Regression
OA	Oleic Acid
OLS	Ordinary Least Square
PARAFAC	PARAllel FACtor analysis

PCs	Principal Components
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Square
PSE	Process Systems Engineering
PMMA	Poly Methyl MethAcrylate
PP	Pour Point
PRESS	Predicted Residual Sum of Squares
PSE	Process Systems Engineering
PY	Pyrogallol
QM	Quantum Mechanics
QSAR	Quantitative Structure Attribute Relationships
QSPR	Quantitative Structure Property Relationships
RMSECV	Root Mean Square Error of Cross-Validation
RMSEP	Root Mean Square Error of Prediction
SA	Stearic Acid
SVM	Support Vector Machines
TBHQ	T-Butyl Hydro Quinone
TG	Triglyceride
TI	Topological Indices
TM	Tensile Modulus
TS	Tensile Strength
UFF	Unified Force Field
XRD	X-Ray Diffraction

CHAPTER 1

INTRODUCTION

1.1 Chemical Product Design

Chemical product design is a response to major changes in the chemical industry which have occurred in recent decades. Traditionally, the chemical industry has focused on process design involving the manufacture of bulk commodity chemicals with an objective primarily being efficient production to reduce costs in order to be competitive. Such chemicals are produced in dedicated equipment and at very large scale such as ethylene, ammonia, and sulfuric acid. The market success of such products depends on the cost of making them. However, in recent decades, specialty, higher value-added, smaller volume chemical products such as pharmaceuticals, electronics, and pigments have become increasingly important. The success for such products depends on their discovery and their time to market. In the production of specialty chemicals, the key is improvement in performance rather than minimization of cost.

Unlike conventional *process systems engineering* (PSE) approaches that were focused on the synthesis, design, optimization and control of chemical processes which are based on *a priori* knowledge on the products, in chemical product design, the identity of the final product is not known. Instead, the basic idea of its behavior is known and the problem is to find the most appropriate chemical(s) that will exhibit and/or cause the desired behavior [1]. Several common desirable behaviors are: biodegradability, lower toxicity, environmental benignity, and less hazard. Since customer needs and customer attributes are the most important sources of product requirements,

product design should address these needs by translating them into the final commercial products. Rapidly developing new products required by customers with minimum cost is an increasingly important problem. According to Cussler and Moggridge [2], chemical product design is a procedure consisting of four stages:

1. **Needs**: Identification of customer needs and the translation of these needs into product specifications.
2. **Ideas**: Generate ideas that satisfy this need.
3. **Selection**: Screen and select the best idea for manufacture.
4. **Manufacture**: Decide what the product should look like and how it should be manufactured.

The 2nd and 3rd stages together represent Molecular Design and Mixture/Blend Design problems. The 1st and the 4th stages may be considered problem formulation and process design stages respectively. Figure 1.1 is a schematic representation of the stages involved in chemical product design and problems that are addressed in each stage [1].

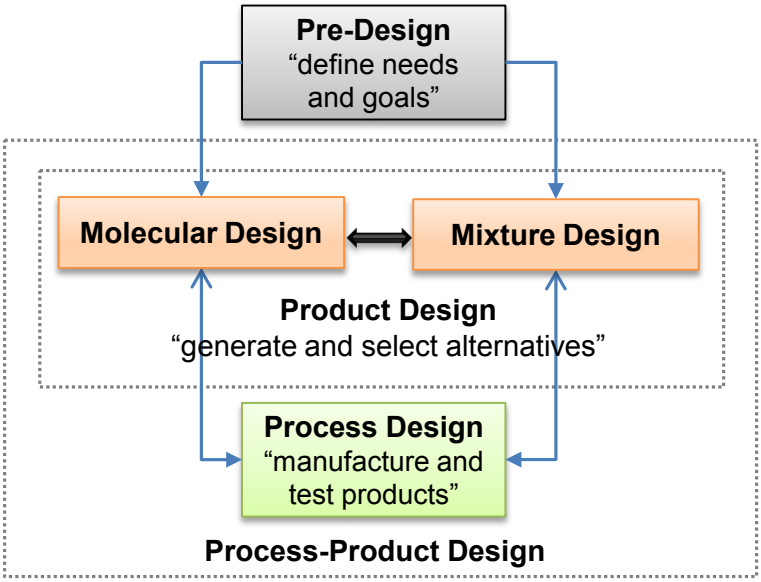


Figure 1.1: Stages involved in chemical product design.

The molecular and mixture/blend design problems can be solved independent of the process design problem or as an integrated product-process design problem. In product design, we describe which product to make; in process design, we explore how we will make it. Chemical product design is the larger topic, and includes the process design in the final step.

Product design is now considered as an emerging paradigm in the field of chemical engineering [1, 2, 3] because it requires a different set of tools and skill sets from other problems traditionally encountered in the field. Design of environmentally benign solvents and alternative media for extraction and purification are new challenges within product design. Examples of chemically formulated products include performance chemicals, paints, cosmetics, pharmaceuticals, proteins, semi-conductors, foods, fuel mixtures, and many more (Figure 1.2).

Figure 1.2: Chemically formulated products.

The search for potentially new molecules that possess one or more desired properties using traditional experimental approaches costs time and resources. This awareness has generated momentum to find alternative ways of numerical characterization of chemicals. Ertl [4] estimated that the

chemical space could be 10^{20} - 10^{24} compounds whereas the current size of the CAS Registry is 2 x 10^{7} compounds. There are abundantly more chemical compounds waiting to be prepared and characterized. Here, the application of computational tools can facilitate the investigation of chemical product formulations prior to experimentation and simulation of their manufacturing processes, giving the flexibility to handle changing design constraints early in the development process. *In silico* research and practice can alleviate problems arising from diverse customer demands and shorter life time of specialty chemical products. When appropriate property models are available to describe and predict the target product properties, computer aided methodologies can be utilized to solve the design problems, and are considered *computer aided molecular/mixture design* (CAMD) problems [1]. Chemical product design that utilizes systematic approaches to integrate the developed chemistry with manufacturing constraints is considered a new chemical engineering paradigm [3, 5].

1.2 Challenges and motivations

Unlike the design of commodity chemicals, which have a known molecular architecture and limited raw material options from which to build an optimum process configuration, the design of specialty chemical products does not have defined molecular architectures or raw material sources. Developing techniques to discover and optimize the molecular architecture that delivers desired attributes is the focus of research in the process systems engineering (PSE) community. In addition, it is important to develop a systematic methodology to produce chemicals that possess both the consumer specified attributes and environmentally acceptable characteristics.

CAMD facilitates the application of computer algorithms to solve the mathematical formulations of chemical product design. For computational searching, the desired property must be calculated from a model describing

the structure-property relationship. Based on the type of the property model, the resulting CAMD problem can be a linear program (LP) or a nonlinear program (NLP). Various CAMD-based techniques have been proposed and can be classified as (a) mathematical optimization techniques [6, 7, 8], (b) stochastic optimization techniques [9], and (c) exhaustive enumeration (generate and test) [1, 10].

Traditionally, molecular and/or mixture/blend design (chemical product design) procedures suffer from two major challenges: the ability to solve large scale optimization problems, and the ability to predict the physical and chemical properties of a given molecule [11]. Computational methods aimed at finding new molecular structures that possess desired product properties follow the experimental approach of *generate-and-test*. These approaches involve the minimization or maximization of an objective function with many linear and non-linear constraint equations. They are commonly referred to as a *Mixed-Integer Non-Linear Programs* (MINLP) [1]. The generic mathematical programming formulation of integrated process-product synthesis/design problems can be presented as [1]:

$$F_{obj} = \max\left\{C^T y + f(x)\right\} \quad \text{Objective function} \tag{1.1}$$

s.t.

$$h_1(x) = 0 \quad \text{Process design specifications} \tag{1.2}$$

$$h_2(x) = 0 \quad \text{Process model equations} \tag{1.3}$$

$$h_3(x, y) = 0 \quad \text{CAMD specifications} \tag{1.4}$$

$$l_1 \le g_1(x) \le u_1 \quad \text{Process design constraints} \tag{1.5}$$

$$l_2 \le g_2(x, y) \le u_2 \quad \text{CAMD constraints} \tag{1.6}$$

$$l_3 \leq By + Cx \leq u_3 \qquad \text{Logical constraints} \qquad \textbf{(1.7)}$$

In the above equations, x represents continuous variables (such as temperature, flowrate and mixture compositions, etc.) and y represents binary integer variables (such as the existence or absence of certain process units, raw materials, or molecular groups). Many variations of the above mathematical formulation exist for solving different chemical product design problems. An in-depth review of the problem formulations has been presented by Gani [1]. The presence of non-linear constraints means that the solution to the MINLP may only be considered a local optimum and not necessarily a global optimum and as such may not be able to generate a complete set of candidate molecules. This leads to sub-optimal design and inefficient iterative solution approach.

Although mathematical programming and hybrid methods provide the framework to transform solution techniques into computer-aided methods and tools to determine the optimal design, they are computationally intensive if the problem is not well-defined or is highly nonlinear. The major non-linearity of the problem is introduced by the property descriptors which either need to be supplied (measured or retrieved from database) and/or predicted through appropriate models. In addition, when considering interfacing product and process design, most algorithms face a bottleneck when it comes to using property models suitable for both product design and process design [1]. If multiple models with restricted application ranges are included for the same property, the algorithm may suffer from discontinuities in the solution trajectory, which may make it more complicated to achieve solution convergence [1]. Overall, purely mathematical optimization-based approaches are computationally intensive and require very detailed models for all considered unit operations, thus it is desirable to reduce the search space prior to invoking the optimization solver.

Almost all CAMD methods have used group contribution-based property prediction methods to evaluate the generated compound with respect to a specified set of desired properties [12, 13, 14]. *Group contribution methods* (GCM) are simple, have acceptable accuracy for many properties and are predictive in nature. However, GCM do not exist for all necessary properties, and reliability of predictions is often questionable for large, complex molecules. There are many property parameters or consumer attributes for which group contribution (GC) data is not available, for example, attributes such as paper softness are difficult to define in terms of properties and molecular architecture.

In order to characterize the properties of observations, one measures variables. In chemical and engineering practice, it is often assumed that our systems are driven by inherent, latent variables. Changes in intrinsic molecular properties cannot easily be measured directly. However, macroscopic properties, which are manifestations of the intrinsic properties, can be predicted through the use of relevant molecular descriptors. Such descriptors can be related to chemical (reactivity, pH, enthalpy of formation, lipophilicity, etc.), physical (boiling point, viscosity, density, etc.), and structural (electron distribution, van der Waals interactions, hydrogen bonds, etc.) properties of molecules. Discrete change of a substituent in one part of a molecule or polymer does not just affect one isolated property, but several of the above properties. Therefore, the chemical behavior of a molecule is governed by these intrinsic, latent molecular properties [15]. Thus, it is important to capture the intrinsic properties using relevant molecular descriptor variables.

Chemical structures can be numerically encoded by molecular descriptors. Thousands of molecular descriptors have been reported in the literature, ranging from simple topological to more complicated topographical properties of molecules. They cover features from constitution to geometry and electronic properties [16, 17]. Relating chemical structures to physical

and chemical properties is not a new endeavor. However, to deal with the enormous amount of data generated, it becomes crucial to develop and utilize features selection algorithms, along with data visualization, interpretation, and mining techniques to progress in computer-aided molecular design.

Molecular descriptor variables can be of spectral origin (e.g. IR, NIR, NMR, UV, X-ray, etc.); chromatographic origin (e.g. HPLC, GC, TLC, etc.) or they may be measurements from sensors in a process (e.g. temperatures, flows, pressure, etc.). When variables are highly correlated or are highly redundant, they are said to be collinear.

As the chemical engineering discipline moves from *data poor* to *data rich*, the research in process systems engineering (PSE) shifts the focus from simulation and complex experimentation toward data-driven techniques to acquire the needed translation of product attributes into common physical-chemical properties [18]. With data-driven means gaining popularity as being powerful and inexpensive, computer usage is being increasingly exploited for scientific investigation. Also, the application of chemometrics and chemoinformatics is gaining a great deal of recognition and application in order to address problems in chemistry, chemical engineering, biochemistry, biology, and medicine [16]. According to Wold [19], chemometrics is the science of extracting chemically relevant information by data-driven means and deriving the respective multivariate statistical models and descriptors. Chemometrics is therefore a process. Chemoinformatics is a subfield of chemometrics which involves the transformation of data into information and information into knowledge to facilitate decision making [18, 16].

Combination of chemometrics and chemoinformatics can lead to target-focused two-and three-dimensional structures from molecular composition. The algorithm and informatics methods can be applicable and transferable to a wide range of chemical and biological systems, regardless of whether the starting point is DNA or a chemical element distribution as illustrated in Figure 1.3 [19].

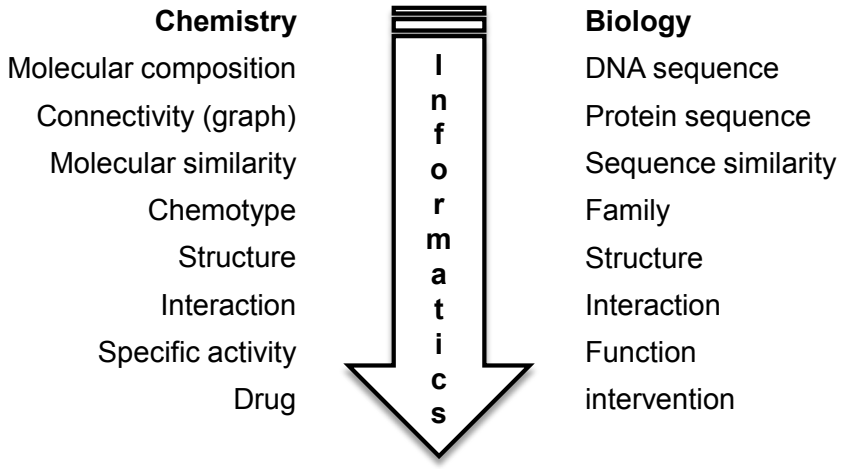


Figure 1.3: Hierarchy of chemical and biological informatics.

Chemometrics can be applied to solve both descriptive and predictive problems involved with such data [19]. In descriptive applications, properties of chemical systems are modeled with the intent of learning the underlying relationships and structure of the system. In predictive applications, properties of chemical systems are modeled with intent of predicting new property values or behaviors of interest. Chemoinformatics can be utilized to analyze a large volume of data generated from molecular modeling, chemical information, and computational chemistry techniques [16].

Moreover, when a good theoretical process model is not available or before new experiments are conducted, historical process data is often available in industry that encompasses a wide spectrum of operating conditions and product grades. Such multi-block data involving the properties of the pure materials, their blend ratios, and the process conditions, could provide an opportunity to enhance the performance of the final product in blending operations for coatings, food, cosmetics, pharmaceuticals and fuels [20]. The data-driven modeling approach is apt for processes and phenomena where cause-and-effect cannot be simply approached from first-principles. For example, a real-life industrial process system often involves complex, nonlinear, incomplete, and uncertain data.

In chemical product (molecular or mixture) design, the identity of the final product is unknown, however, the general behavior or characteristics of the product (goal) is known. Since the properties of the component or mixture of components dictate whether or not the design is useful, the basis for solution approaches in this area should be based on the properties themselves. The recognition of property-based design came about as a direct result of the following observations:

- Many processes are driven by properties not components.
- Performance objectives are often described in terms of measurable physical-chemical properties.
- Often objectives cannot be described by composition alone.
- Molecular/mixture design is based on properties.
- Insights are often hidden by not integrating properties directly.

The discussion presented here provided the motivation that guided the research, and as a result the methods and tools developed must accomplish the following:

- Integrate process and product design problems via a methodology within the property integration paradigm.
- The approach needs to be systematic and capable of setting up the design performance requirements or "targets" *a priori*, i.e. a targeting approach.
- Utilize a data-driven approach combined with multivariate statistical techniques to solve both descriptive and predictive problems.
- Apply multivariate characterization, modeling, and design.
- Incorporate the concepts of reverse problem formulation and property clusters to aid in the decomposition of the design problem.
- The technique should take advantage of the benefits of visualization tools in the formulation of the problem and as part of its solution algorithms.

1.3 Scope and Objectives

General background information on traditional frameworks for solving computer-aided molecular/mixture design (CAMD) problems is presented in this work. Techniques such as group contribution, reverse problem formulation, property clustering, and multivariate statistics are presented within CAMD to formulate and solve chemical product design problems. These methods form the basis for the property-based design techniques presented in this dissertation.

Primarily, the effort of this dissertation is on the formulation of chemical products through the application of multivariate characterization, modeling, and design. Property-based molecular and mixture design algorithm within the general property clustering framework utilizing chemometric techniques and group contribution methods in a reverse problem formulation are presented. Data-based modeling and optimization strategies will be presented in this dissertation covering two specific areas of chemical product design:

1. **Mixture/Blend design**: For mixture products, a systematic and insightful use of past data describing the properties of the raw materials, their blend ratios, and the process conditions during the production of a range of product grades is investigated to explore their effects on the final product properties and to achieve new and improved products.
2. **Molecular design**: For molecular products, chemically feasible, structured molecules are exhaustively enumerated from a set of appropriate descriptors (represented by fragments or building blocks) to identify compounds exhibiting certain desirable or specified behavior.

The present work is limited to systems that can be characterized by three properties, however, algebraic and optimization-based approaches are available to extend the application range to include more properties [21].

1.4 Significance of the Research

A significant result of the developed methodology is that for problems that can be adequately described by just three properties, the process and molecular design problems are solved visually and simultaneously on a ternary diagram, irrespective of how many substructural molecular fragments are included in the search space. The primary benefits gained through the developed methodology are as follows:

- utilization of interdisciplinary methods and tools that extend through multivariate statistics, applied mathematics, computational expertise, and experimentation across all scales,
- formulation and solution of chemical design problem on property basis,
- bridging domains via property models facilitate nested and nonlinear routines to be reduced to a single sub-problem in cluster space.
- identification of design targets without performing detailed calculations,
- visualization and solution of problem in lower dimensional space, which provide valuable information by elucidating hidden relationships in data and multivariate understanding of complex processes and phenomena,
- multivariate projection-based regression retains better memory of the structure of the training set data to predict molecules with similar properties and structures,
- reverse design approach using the descriptors of group contribution type combined with chemoinformatics enables exhaustive search of the structures corresponding the target physical properties, and
- explicitly tracking and integrating properties in a systematic manner relieve, the iterative nature, multiple component combinatorial explosion, and difficulty of formulating and solving the mixed integer non-linear programs (MINLP) of conventional design techniques.

The aim is to allow us to see the *big picture first, and the details later*.

1.5 Organization

Chapter 2 introduces the general theoretical background on the methods and tools that the research in this dissertation is based on. Section 2.2 introduces the property-based approach that facilitates the flow of information from the process level to the molecular level, and vice versa. Section 2.3 highlights property estimation methods based on combinatorial techniques such as the group contribution method. Different characterization (e.g. IR spectroscopy) based group contribution methods are presented in Section 2.4. The characterization-based groups serve as the descriptive application of chemometric techniques that describe the molecular architecture of a chemical product. Section 2.5 discusses the limitation of using general regression models that is often encountered when the number of descriptor variables is larger than the number of samples and when the variables exhibit linear relationships with each other. The predictive application of chemometric techniques is presented using statistical multivariate techniques like PCA and PLS to model the properties of chemical systems in Section 2.6.

Chapter 3 presents the developed methodology for solution of mixture design problems utilizing the tools presented in Chapter 2. The multi-block data structures (L- and T-shaped) available in blending operations are presented in Section 3.3. Combining multiple blocks using appropriate mixing rules and matrix algebra to help simplify the analysis and design by not having to differentiate between mixture and process variables and not having to assume independence of the factors when multivariate analysis techniques are used is discussed in Section 3.4. Section 3.5 presents a case study to illustrate the method and concept using the development of thermo-plastic as a case study. The example incorporates all three degrees of freedom available in blending operation.

Chapter 4 presents the developed molecular design framework based on a characterization-based group contribution method. The advantages of

biodiesel, a short overview on its production, chemical profiles of different feedstocks and their effects on product quality are present in Section 4.1. Technical difficulties with biodiesel that have impaired its use and commercialization are presented in Section 4.3. Additive design is presented as an alternative approach to solve the technical difficulties with biodiesel in Section 4.4. Section 4.5 discusses the methodology for additive design that combines the framework and the tools presented in Chapter 2. In this Section, the types of training set additive molecules, their property estimation methods, IR-based characterization of their molecular architecture, multivariate statistical analysis of IR data for descriptive application, latent variable property model development for predictive application, evaluation of target properties for additive design, and design of additive molecules that meet the target property specifications are discussed.

Chapter 5 introduces the use of density functional theory (DFT) based simulation techniques to generate the IR spectra as molecular descriptors with which to develop predictive property models. This step relieves the dependency on measured or database values for the required IR data used in the methodology presented in Chapter 4. The interdisciplinary and novel framework proposed will be demonstrated using a case study focused on the reverse design of ionic liquids with tailored properties. The methods in this Chapter will integrate the framework and tools presented in Chapter 2 and Chapter 4 with an additional ability to generate spectroscopic data from quantum chemical calculations.

Chapter 6 introduces future works that can be extended from the methodologies and tools developed in this dissertation. Section 6.1 introduces potential improvement or extension areas in the characterization, modeling, and design methodologies presented in this dissertation. Section 6.2 incorporates the design of inherently benign chemical process routes using process route descriptors related to environment, health, and safety factors.

CHAPTER 2

THEORY AND METHODOLOGY

In this chapter, theoretical background on current methods used in chemical product design along with tools that will be employed in this dissertation research will be defined and discussed.

2.1 Process-Product Design

The molecular and mixture/blend design problems can be solved independent of the process design problem or as an integrated product-process design problem as shown in Figure 2.1. In general, the objective in the design or optimization of process is to find a balance between satisfying process unit requirements/constraints and the use of appropriate raw materials in order to maximize profit or to minimize cost. Traditionally, process design and molecular/mixture design have been treated as two separate problems, with little or no feedback between the two approaches as represented by Figure 2.1 [22]. The raw materials could be searched from a material database or can be obtained from molecular design. Since process and molecular/product design are decoupled, molecular design can be performed based on qualitative process knowledge and/or experience. This can lead to sub-optimal designs and an inefficient iterative solution approach.

However, the decoupling of the process and the molecular design problems can be addressed by systematically solving a series of reverse problems [22, 24]. This is accomplished by a property-based approach which

facilitates the flow of information from the process level to the molecular level, and vice versa.

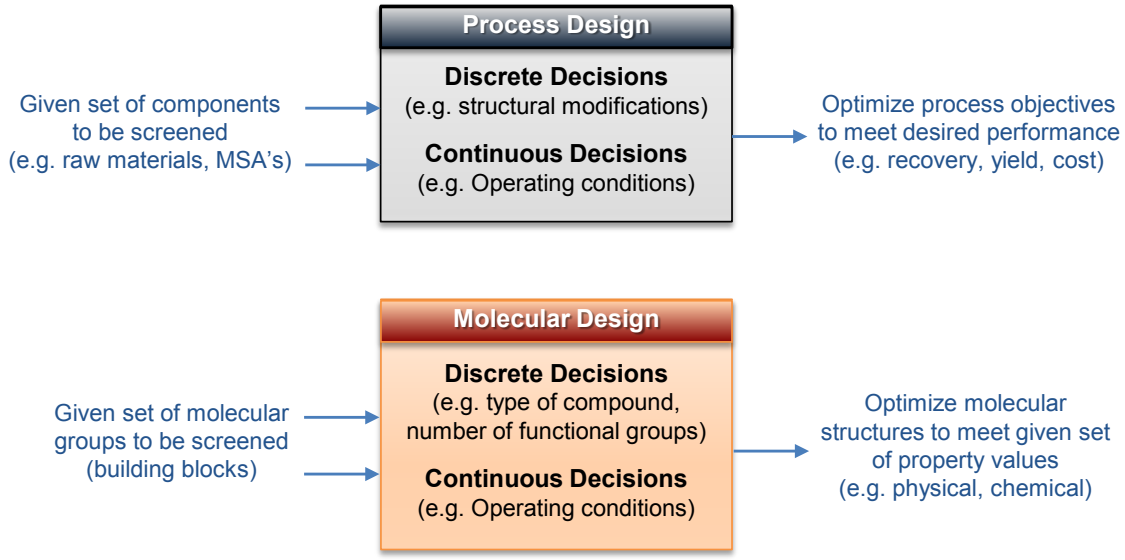


Figure 2.1: Conventional solution approach for process and molecular design problems.

2.2 Property-Based Process and Product Design

In molecular or mixture design, the identity of the final product is unknown, however, the general behavior or characteristics of the product (goal) are known. The objective is to find the most appropriate chemical, or mixture of chemicals, that will satisfy these goals. Since the properties of the component or mixture of components dictate whether or not the design is useful, the basis for solution approaches in this area should be based on the properties themselves. Similarly, in process design, since properties (or functionalities) form the basis of performance of many process units (e.g., vapor pressure in condensers; specific gravity in decantation; relative volatility in distillation; Henry's coefficient in absorption; density and head in pumps; density, pressure ratio, and heat capacity ratio in compressor; etc.), it would be very insightful to develop procedures based on key properties instead of key components [22].

However, unlike component-based *chemo-centric* approaches where chemical components are conserved and mixing of components is linear; in a *property-based* approach, properties are not conserved entities and mixing of properties is not necessarily linear. To overcome these limitations, Shelley and El-Halwagi *et al.* [23] introduced the property clustering framework that uses conserved quantities called *clusters*. It was later applied to process and product design by Eden *et al.* [22]. To bridge the gap between the process design and the molecular design problems, Eljack *et al.* [24] extended the property integration framework by combining the property clustering technique and Group Contribution Methods (GCM). Qin *et al.* [21] introduced an algebraic approach using property clusters to relieve the limitation in the use of more than three properties. These contributions enabled simultaneous consideration of process performance requirements and molecular property constraints in the cluster domain and solution of the design problem in this reduced domain. The following sections discuss the property integration framework based on reverse problem formulation and property clustering techniques.

2.2.1 Reverse Problem Formulation

The reverse problem formulation technique decomposes a process-product design problem into two reverse problems linked by property targets/constraints [22, 24]. This enables a two-step approach, where the property targets that satisfy the process performance/constraints are identified in the first step and then the molecules that match the targets are identified in the second step. This gives the ability to identify optimum solutions to process and product design problems more easily than solving the conventional forward problems, which are iterative in nature [22, 24, 25, 27].

Figure 2.2 is a schematic representation of the reverse problem formulation concept. The first reverse problem is the reverse of a simulation problem, where the process model is solved in terms of the constitutive/design

variables instead of the process variables, thus providing the design targets. The second reverse problem solves the constitutive equations (property models) to identity candidate components by matching the design targets.

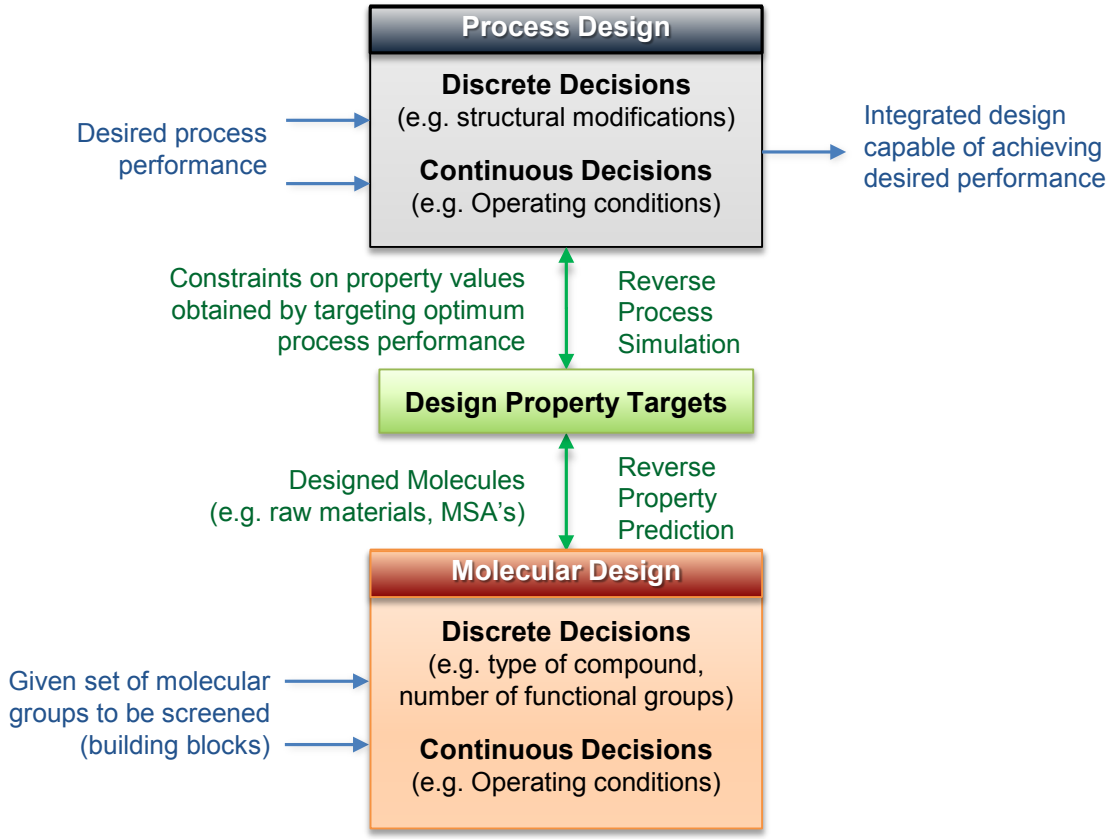


Figure 2.2: Product and process design problem using reverse problem formulation methodology.

Prediction of macroscopic properties from molecular information is nontrivial and is by no means extensive. In most methods, the computed information at smaller length scales is passed to models at a larger length scale. This is achieved by coarse-graining (removing degrees of freedom) and information passing between low- and high-scales [26]. Pathway I in Figure 1 represents the dependency of overall higher scale performance on the lower scale phenomena. This path (approach) necessitates coupling of mathematically different models and phenomena across two or more scales.

This results in a greater coupling in design of products at multiple scales, thereby increasing problem complexity, and combinatorial explosion in the number of models and parameters [12].

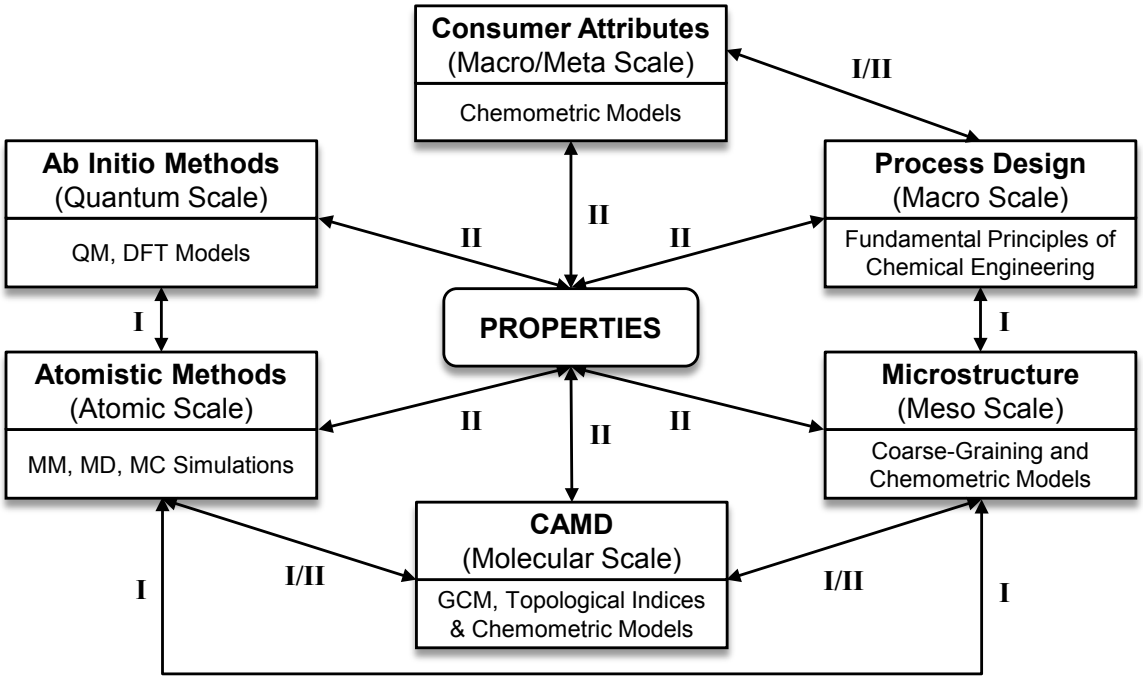


Figure 2.3: A multi-scale product design framework showing (I) traditional approach and (II) RPF approach linking each of the scales via a common property domain.

The use of reverse problem formulation (RPF) helps circumvent the challenge posed by coupling of scales by bridging them through a property domain (pathway II in Figure 2.3) [27]. RPF uses the duality of linear programming to reformulate the design problem as a series of reverse problems solved in the property domain. This way, an immense computational cost associated with the hierarchical nesting across multiple-scales is relieved leading to a much more efficient solution achieved through reduction in the need for enumeration [27].

2.2.2 Property Clustering Technique

Property clustering is a property-based visualization tool for mapping the design problem from the non-conserved property domain into the conserved (component-less) cluster domain [23]. The property clustering technique utilizes property operators, which are functional relationships describing the attributes and physical-chemical properties. The clusters map property relationships into a lower dimensional domain, thus allowing for visualization and insights into the problem. Details can be found in Eden *et al*. [22]. Only highlights will be presented in this dissertation.

2.2.2.1 Property Operator Functions

The clustering approach utilizes property operators, Ψ, which are functions of the original raw physical properties. In Eq. (2.1), the property is described by a general linear mixing rule:

$$\psi_j\left(y_j\right)_{Mix} = \sum_{i=1}^{n} x_i \cdot \psi_j\left(y_j\right)_i \qquad \textbf{(2.1)}$$

where, x_i = the fractional contribution of component i

y_j = the j^{th} property

Although the property operator equation must have linear mixing rules, the property operator itself may be nonlinear. For example, if the property operator describes density, then to meet the linear criteria imposed by Eq. (2.1) we would use specific volume as the property operator, ignoring any interaction effects from mixing, as shown in Eq. (2.2) and (2.3). The operator expressions will invariably be different for molecular fragments and process streams, however, as they both represent the same property, they can be visualized on a common cluster domain in a similar fashion.

$$\psi_j\left(y_j\right)_i = \frac{1}{\rho_i} \tag{2.2}$$

$$\frac{1}{\rho_{Mix}} = \sum_{i=1}^{n} x_i \cdot \frac{1}{\rho_i} \tag{2.3}$$

Since the properties may have various functional forms and units, the operators are normalized into a dimensionless form by dividing by an arbitrary reference operator and then summarized to yield an Augmented Property Index (*AUP*) as:

$$\Omega_{ji} = \frac{\psi_j\left(y_j\right)_i}{\psi_j\left(y_j\right)_{ref}} \quad , \quad AUP_i = \sum_{j=1}^{p} \Omega_{ji} \tag{2.4}$$

A cluster is then defined by dividing the non-dimensionalized property by the *AUP*, as:

$$C_{ji} = \frac{\Omega_{ji}}{AUP_i} \quad \text{,where } \sum_{i=1}^{k} C_{ji} = 1 \tag{2.5}$$

Since the clusters are tailored to maintain the two fundamental rules for *intra-* and *inter- stream* conservation, additive rules, e.g. *lever-arm rules*, are needed to ensure that the mixture property cluster of two streams with different individual property clusters can be easily determined. The mixture cluster, Eq. (2.6), and mixture *AUP,* Eq. (2.7)*,* values can be calculated through the linear mixing rules as follows:

$$C_{j,Mix} = \sum_{i=1}^{n} \beta_i \cdot C_{ji} \quad \text{, where } \beta_i = \frac{x_i \cdot AUP_i}{AUP_{Mix}} \tag{2.6}$$

$$AUP_{Mix} = \sum_{i=1}^{n} x_i \cdot AUP_i \tag{2.7}$$

The relative cluster arm, β_i, which represents the fractional contributions of the streams, can be calculated using Cartesian coordinates as well as estimated visually from the relative lengths of the lever arms as shown in Figure 2.5. Visually, intra-stream conservation means that once two clusters are known, the third one is automatically determined. In order to use common tools like Microsoft Excel® that do not support ternary plot representation, the ternary coordinate is converted to Cartesian coordinates.

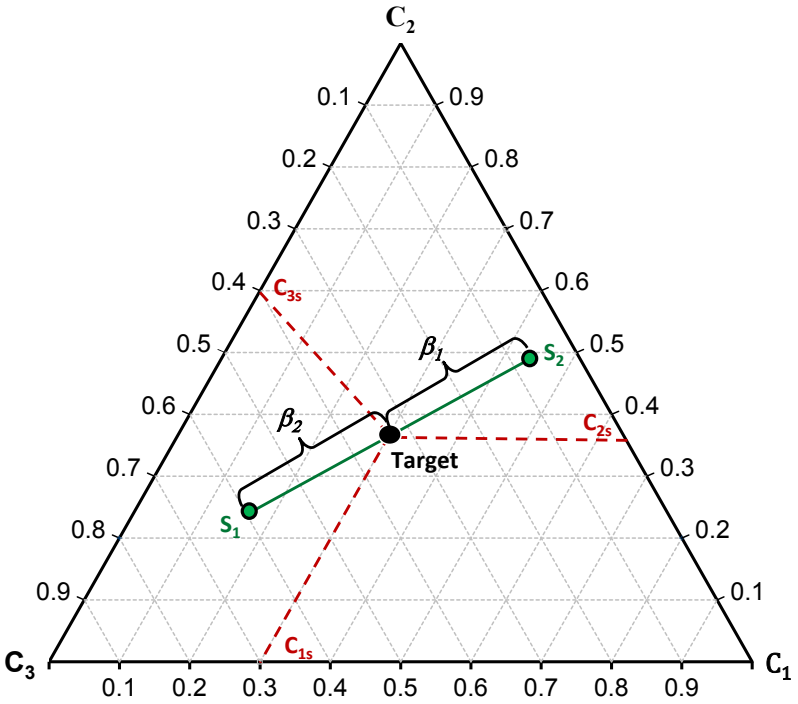


Figure 2.4: Representation of intra- and inter-stream conservation of clusters in ternary diagram.

2.2.2.2 Ternary to Cartesian Coordinates

The coordinate transformations from ternary to Cartesian for an equilateral triangle can be accomplished using the *Pythagorian theorem* in terms of ternary coordinates defined as cluster values and dimensionless property operators are as follows:

$$X_{CC,s} = C_{1s} + 0.5C_{2s} = \frac{\Omega_{1s} + 0.5 \cdot \Omega_{2s}}{\Omega_{1s} + \Omega_{2s} + \Omega_{3s}} \quad \textbf{(2.8)}$$

$$Y_{CC,s} = C_{2s} = \frac{\Omega_{2s}}{\Omega_{1s} + \Omega_{2s} + \Omega_{3s}} \quad \textbf{(2.9)}$$

Furthermore, the relative arms can be calculated accordingly as:

$$\beta_1 = \sqrt{\frac{(X_{CC,2} - X_{CC,mix})^2 + (Y_{CC,2} - Y_{CC,mix})^2}{(X_{CC,2} - X_{CC,1})^2 + (Y_{CC,2} - Y_{CC,1})^2}} \quad \textbf{(2.10)}$$

$$\beta_2 = \sqrt{\frac{(X_{CC,1} - X_{CC,mix})^2 + (Y_{CC,1} - Y_{CC,mix})^2}{(X_{CC,2} - X_{CC,1})^2 + (Y_{CC,2} - Y_{CC,1})^2}} \quad \textbf{(2.11)}$$

In cluster space, points are used to represent discrete property values while feasibility regions are used to represent a range of accepted property values (Figure 2.5).

2.2.2.3 Feasibility Region Boundaries

For problems that can be adequately described by just three properties, the desired process performance range or product property requirement range can be visualized within the boundary of the true feasibility region defined by six unique points, Eq. (2.12), characterized by values of dimensionless operators. The feasibility region narrows down the search space and guarantees that no feasible points will exist outside it. The feasibility region boundary analysis helps identify and describe the expression for the feasibility region *a priori* and requires no enumeration of an infinite number of feasible points possible within its boundaries [22, 23]. In Figure 2.5, the feasibility region is shown in dashed lines, to form a hexagon in a ternary cluster diagram.

$$\begin{matrix} \left(\Omega_1^{min},\Omega_2^{min},\Omega_3^{max}\right) & \left(\Omega_1^{min},\Omega_2^{max},\Omega_3^{max}\right) & \left(\Omega_1^{min},\Omega_2^{max},\Omega_3^{min}\right) \\ \left(\Omega_1^{max},\Omega_2^{max},\Omega_3^{min}\right) & \left(\Omega_1^{max},\Omega_2^{min},\Omega_3^{max}\right) & \left(\Omega_1^{max},\Omega_2^{min},\Omega_3^{min}\right) \end{matrix} \quad \textbf{(2.12)}$$

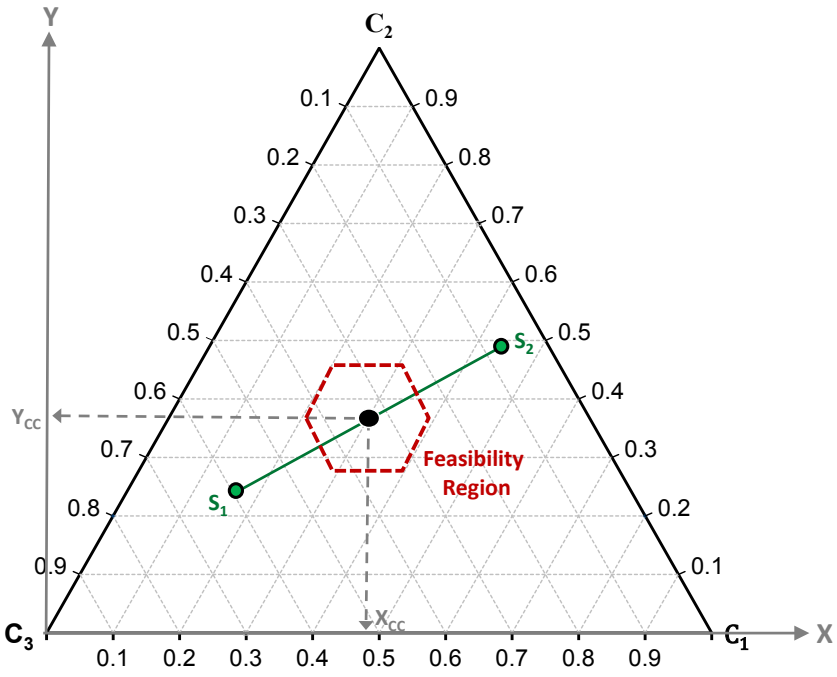


Figure 2.5: Representation of feasibility reason with source-sink mapping using clusters in ternary diagram. The clustering points are converted from ternary to Cartesian coordinate.

Visualization of the problem allows for easy identification of optimum strategies for the combination of molecular groups, while the unique feature of linear mixing rules allow for the use of simple lever arm analysis to solve the problem in reduced cluster space. This way, the design problem can be solved by identifying the product properties corresponding to the desired process performance. The property clustering methods for solving mixture and molecular design problems form the basis for the methods developed in this dissertation.

Design of target chemical products that have tailored structures and that exhibit an array of unique functional properties requires the use of

physical-chemical properties and/or biological activities. Experimental measurements of properties of interest for products such as drugs, ionic liquids, additives, etc. are scarce and limited. With respect to the synthesis, characterization, and applications of such chemical products, it is essential to estimate properties through theoretical or empirical means.

2.3 Prediction of Properties

As process-product design may involve both process and molecular design, separate property models may be required. Property models can range from easy to use data-dependent regressed models whose application range is small, to computationally complex models like *ab initio* (quantum-chemical) calculation, which can be applied to any chemical system. In the middle are the most common methods that balance full empiricism with basic chemical theory to adapt the rigorous equations that describe molecular architecture in a form suitable for design. A comprehensive overview of a wide range of property estimation methods can be found in Poling *et al.* [28]. Constantinou and Gani [12] classified property estimation methods as shown in Figure 2.6.

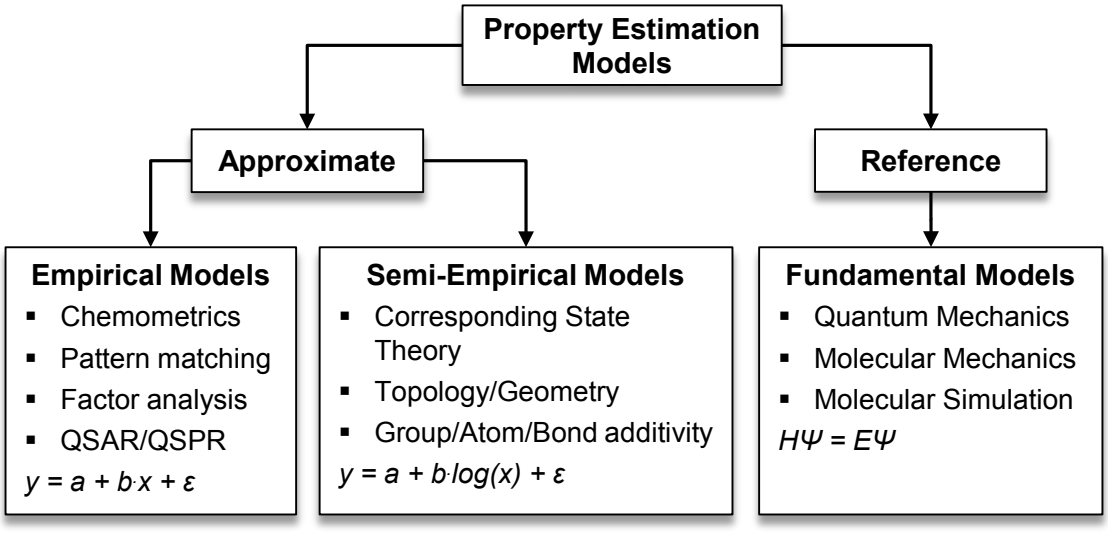


Figure 2.6: Classification of property estimation methods.

Regardless of the type of property model used, it is essential that they are easy-to-use, accurate, reliable, and computationally efficient throughout the entire design domain. *Group contribution* (GC) and *topological indices* (TI) are two such techniques that utilize *quantitative structure property/activity relationships* (QSPR/QSAR) that have been developed using empirical relationships between molecular architecture and physical-chemical properties found in large databases [11, 29, 30].

A QSPR model finds the relationship between structural features (constitutional, topological, geometrical, etc.) and physical, mechanical, or chemical properties of materials. In drug design, the QSAR model finds the underlying relationship between molecular descriptors with pharmacokinetic/pharmacodynamics, and ADMET (Absorption, Distribution, Metabolism, Elimination and Toxicity) properties [16, 31, 32]. Since QSPR models provide information on features affecting the compounds' physicochemical properties, they can be used for screening and further optimization. Thus, development of robust QSPR model can facilitate conserving resources and accelerating the process of development of new and enhanced products.

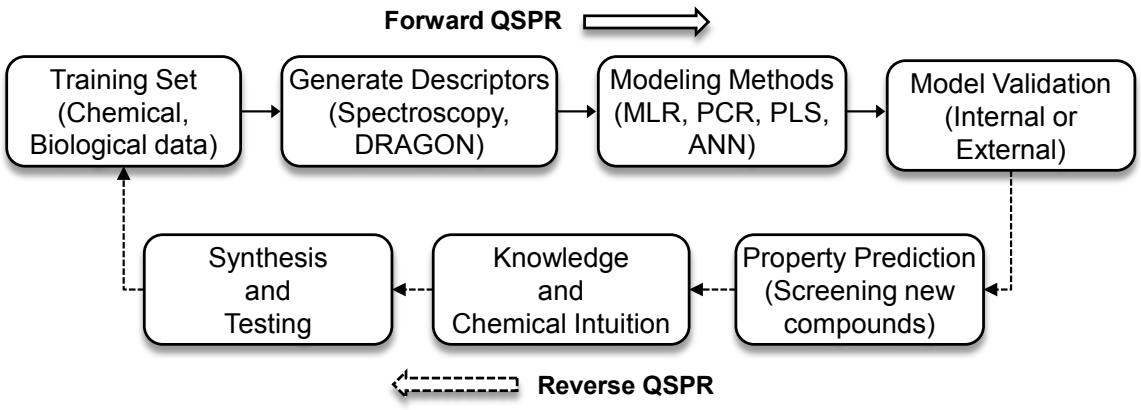


Figure 2.7: QSAR model development steps.

Figure 2.7 presents QSPR model development steps. Several machine learning methods such as *artificial neural network* (ANN), Bayesian statistics, *support vector machines* (SVM), *genetic algorithm* (GA), *partial least-squares analysis* (PLS), and *multiple linear regression analysis* (MLR) have been used in literature to perform QSPR modeling [31].

However, in order for a developed model to be relevant to the target product properties, the training set must contain optimum molecules covering balanced variation (high diversity) of the features spanning the chemical/property space believed to be important for interaction with the physical, chemical, or biological target [30]. The diversity of the chemical features can be achieved by using statistical experimental design such as *factorial design* and *D-optimal design*. The coverage can be achieved by employing quadratic or cubic models instead of linear model design [30, 33]. However, the requirement of high diversity and full coverage often leads to redundancy. Since a set of structural descriptors is never complete, some redundancy could be allowed at this level. The use of projection methods (discussed later in Section 2.6) will circumvent the problem of redundancy in the descriptor data.

2.3.1 Group Contribution Method

Most property estimation methods are based on the group contribution method (GCM), where appropriate descriptors or predefined fragments (group, bond, or atom, etc.) representing a molecule are identified and the properties of the molecule are estimated by summing all the contributions from each fragment that make up the molecule [12, 14, 22, 24, 25, 34]. The GCM is a powerful product/molecular design tool, which allows prediction of the physical properties of molecules from structural information alone. An additive, three level, group contribution property estimation model, which estimates the property of a compound as a linear combination of the appropriate descriptor contributions, is as follows:

$$\underbrace{f(x)}_{\text{Property Function}} = \underbrace{\overbrace{\sum_i N_i C_i}^{1^{st}\ \text{Order}} + \overbrace{\sum_j M_j D_j}^{2^{nd}\ \text{Order}} + \overbrace{\sum_k O_k E_k}^{3^{rd}\ \text{Order}}}_{\text{Group Contribution Terms}} \quad \textbf{(2.13)}$$

where, C_i = the contribution from first-order group i

N_i = the number of occurrence of first-order group i

D_j = the contribution from second-order group j

M_j = the number of occurrence of second-order group j

E_k = the contribution from third-order group k

O_k = the number of occurrence of third-order group k

First order groups contain basic information and can be combined linearly since they assume no interaction between groups. Second order groups can be estimated from first order groups and correct for the interactions between first order groups. Third order groups can be derived in a similar manner and help to correct for poly-functional compounds with more than four carbon atoms in the main chain.

However, any application of group contribution relies on the availability of atom type, molecular group type, or type of chemical bonding present to describe the structure as well as tables giving the property contributions of each group. There are many properties which cannot be estimated by GCM. For instance, cetane number is an important performance indicator for biodiesel, but GCM parameters are not available to describe this property. Furthermore, not all possible atomic arrangements and structures can be represented in GCM. Hence, there is a need for an efficient methodology for the design of structured molecules. One such approach to structured product design is combining decomposition techniques with multivariate methods [35, 36, 37].

This approach first utilizes multivariate characterization techniques such as infrared (IR) and near infrared (NIR) spectroscopy to describe a set of representative samples, and then uses decomposition techniques such as *principal component analysis* (PCA) and *partial least squares onto latent surfaces* (PLS) to find the underlying latent variable models that describe the molecule's properties. The factors are called 'latent' because they cannot be observed directly, but can be characterized indirectly. The orthogonal nature of these models allows for group-based interpretations and property predictions which can be utilized to design new molecules not found in the original set of molecules. The structure and identity of candidate molecules can then be identified by combining or "mixing" substructural molecular fragments until the resulting properties match the targets [24].

2.4 Characterization Techniques

Characterization is a class of tools associated with the determination of not only chemical constituents or molecular structure, but also of larger structural characteristics describing the orientation and alignment of these molecules by exploiting the fact that molecules absorb specific frequencies that are characteristic of their structure. Some common characterization techniques include nuclear magnetic resonance (NMR), x-ray diffraction (XRD), and infrared spectroscopy (IR). Characterization techniques are often applied to a training set of molecules used to explore a set of property attributes. Solvason [27] proposed a general guide for managing the complexity of the information through a systematic method for determining which specific information on molecular architecture will be necessary to build appropriate models for a specific application as shown in Figure 2.8.

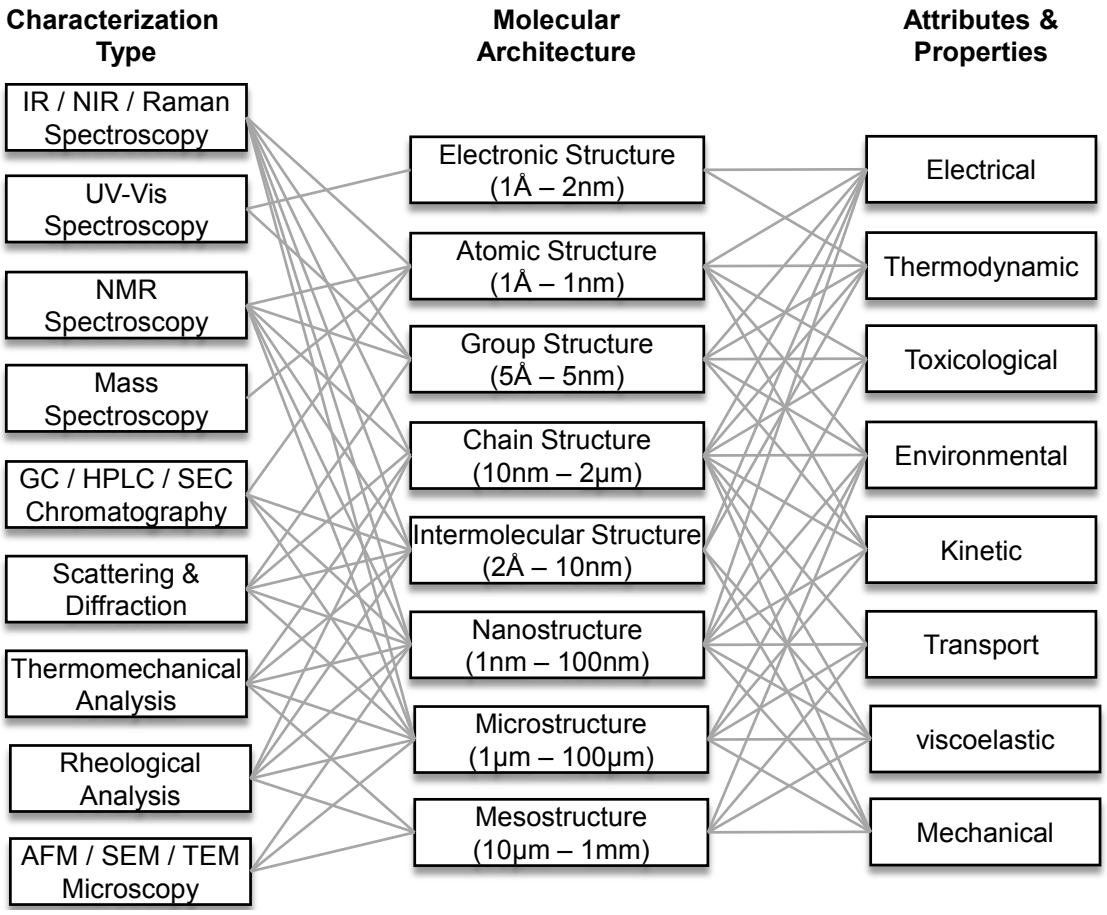


Figure 2.8: An overview of the interconnectivity of characterization techniques, molecular architecture, and physical properties and attributes of chemical and material products.

Spectroscopy will be the primary characterization technique explored in this dissertation. Spectroscopy such as infrared (IR) and near-infrared (NIR) provide specific information on the presence of functional groups, information on the orbital configurations of the electrons, and details of the carbon-hydrogen structure of the chemical products. The added structural information available from this characterization can also be used to distinguish some orientation specific information in various isomer geometries. More on spectroscopy can be found elsewhere [38]. Appendix B gives an overview on IR spectroscopy.

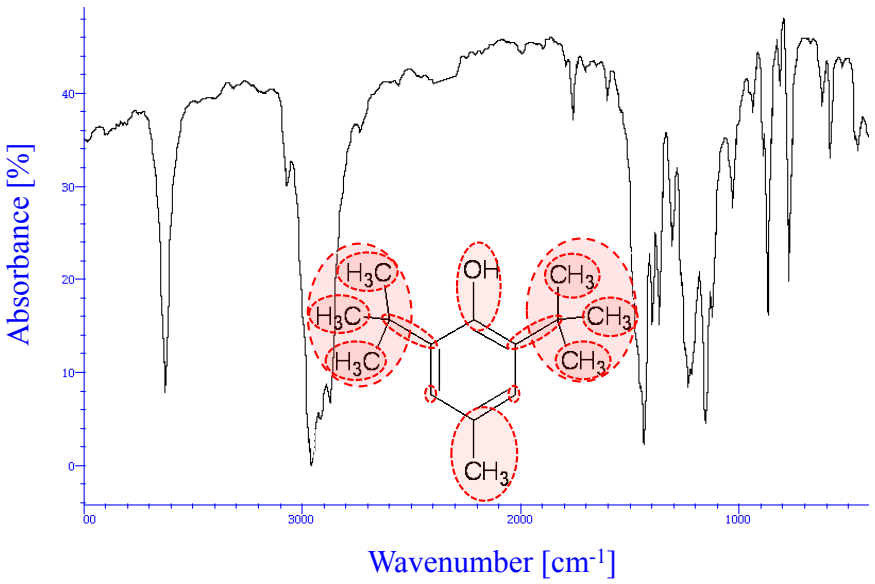


Figure 2.9: IR spectra of butylated hydroxytoluene molecule.

Figure 2.9 is an IR spectrum of a butylated hydroxytoluene molecule [39] and its molecular structure showing 1st and 2nd order GC groups. IR spectra of such molecules contain large quantities of descriptor data involving information on molecular architecture. In addition, there is a high probability that many of the descriptor variables will be correlated, i.e., that some variables will be linear functions of other variables [40]. Managing such complexity of information to design chemical products and to build appropriate models for a specific application will require a systematic method for capturing important features of the molecular architecture. Therefore multivariate statistical techniques can be used to decompose large quantities of information about the system in the initial training set. More detailed discussion on multivariate techniques will be presented in Section 2.6.

2.4.1 Spectroscopy

Many substances in solution follow *Beer-Lambert's law,* showing a linear relationship between concentration and absorbance. Beer's law relates the absorption of light to the properties of the material through which the light is

traveling. The absorbance of a single wavelength due to the presence of a single component can be written as [41]:

$$A = K \cdot C \tag{2.14}$$

where, A = the absorbance at that wavelength

K = the absorbance coefficient for that wavelength and component

C = the concentration of the component

This concept can be applied to any system where we can measure a quantity, A, that is proportional to some property, C, of the sample. The generalization of the above equation for multiple components, multiple wavelengths and multiple samples gives:

$$A_{ws} = \sum_{c=1}^{n} K_{wc} \cdot C_{cs} \tag{2.15}$$

where, A_w = the absorbance at the w^{th} wavelength in the s^{th} sample

K_{wc} = the absorbance coefficient for the w^{th} wavelength and for the c^{th} component

C_c = the concentration of the c^{th} component in the s^{th} sample

n = the total number of components

In expanded form Eq. (2.15) takes the form:

$$\begin{array}{ccccccccc}
A_{1s} & = & K_{11} \cdot C_1 & + & K_{12} \cdot C_2 & + & \ldots & + & K_{1c} \cdot C_{cs} \\
A_{2s} & = & K_{21} \cdot C_1 & + & K_{22} \cdot C_2 & + & \ldots & + & K_{2c} \cdot C_{cs} \\
A_{3s} & = & K_{21} \cdot C_1 & + & K_{32} \cdot C_2 & + & \ldots & + & K_{3c} \cdot C_{cs} \\
\vdots & & \vdots & & \vdots & & \ddots & & \vdots \\
A_{ws} & = & K_{21} \cdot C_1 & + & K_{w2} \cdot C_2 & + & \ldots & + & K_{wc} \cdot C_{cs}
\end{array} \tag{2.16}$$

It is clear from Eq. (2.16) that the absorbance at a given wavelength, w, is simply the sum of the absorbances at that wavelength from each of the components present. Since matrix notation can easily generalize the above Eq. (2.16), a single matrix equation is:

$$\mathbf{A} = \mathbf{K} \cdot \mathbf{C} \qquad \textbf{(2.17)}$$

Note: The notations used here are limited to this section only. New notations will be introduced in following sections that will be used throughout the remainder of the dissertation. Boldfaced unsubscripted letters are used to refer to vectors and matrices.

In spectroscopy, it is clear that the spectrum of a sample is the sum of the spectra of the constituents multiplied by their concentrations in the sample. If the concentrations are t and the spectra p, we get the latent variable model $X = t_1 \cdot p_1^T + t_2 \cdot p_2^T + ...t_c \cdot p_c^T = T \cdot P^T + noise$ (Beer-Lambert's law). In many applications this interpretation with the data explained by a number of "factors" (components) makes sense [41]. Chemometric modeling methodologies can be investigated and employed as a means to derive mathematical relationships between spectroscopic measurements and measured product properties. Kramer [41] suggests that the power and limitations of chemometric techniques that should be realized while solving data-driven problems are as follows:

We can use these chemometric techniques to:

- remove as much noise as possible from the data.
- extract as much information as possible from the data.
- use the information to learn how to make accurate predictions about unknown samples.

In order for this to work, two essential conditions must be met:

- the data must have information content.
- the information in the data must have some relationship with the property(s) which we are trying to predict.

Because consumer attributes are difficult to quantify physically, the relationship between them and the underlying fundamental physical-

chemical properties and/or the molecular architecture will most likely involve empirical relationships. Empirical models describe the underlying phenomena's relationship to a set of experimental data using regression analysis.

2.5 General Regression Models

Traditionally, the modeling of the *predicted response*, **Y**, by means of a *descriptor variable*, **X**, is done using multi-linear regression (MLR), which works well as long as the **X**-variables are fairly few and fairly uncorrelated, i.e., **X** has *full rank* (rank is a number expressing the true underlying dimensionality of a matrix). Often, the relationship between **X** and **Y** variables can be approximated using a linear model and can be represented mathematically as:

$$\mathbf{Y}_{M\times L} = \mathbf{X}_{M\times K}\cdot\boldsymbol{\beta}_{K\times L} + \mathbf{E}_{M\times L} \tag{2.18}$$

where, M = the number rows of sample readings or observations

L = the number columns of measured response properties

K = the number columns of descriptor variables (like components)

$\boldsymbol{\beta}$ = the regression coefficients or sensitivities matrix

$\mathbf{E}$ = the error or residual matrix

Three cases can be distinguished as described by Geladi and Kowalski [42] in Eq. (2.18) to determine $\boldsymbol{\beta}$:

1. $K > M$: There is no unique solution for $\boldsymbol{\beta}$ as infinite numbers of solutions exist, unless one deletes predictor variables.
2. $K = M$: There is one unique solution provided that **X** has full rank.
   $\mathbf{E} = \mathbf{Y} - \mathbf{X}\cdot\boldsymbol{\beta} = 0$
3. $K < M$: There is no exact solution for $\boldsymbol{\beta}$, however, a solution can be achieved by minimizing the residual in the following equation:
   $\mathbf{E} = \mathbf{Y} - \mathbf{X}\cdot\boldsymbol{\beta}$

The *ordinary least-square* (OLS) method is the most popular method to find the regression coefficients by maximizing the model sum of squares and minimizing the residual sum of squares. Using least-square, **β** can be estimated by:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}^T \cdot \mathbf{Y} \qquad \textbf{(2.19)}$$

where, the superscript T symbolizes the transpose of a matrix.

When the number of **X**-variables is large compared to the number of observations, for example in spectroscopy ($i.e. K > M$), it can lead to a singular ($\mathbf{X}^T\mathbf{X}$) matrix whose inverse does not exist. This is because the number of unknown variables is greater than the number of equations, leading to an underdetermined equation system which has an infinite number of solutions for **β**. This is the most frequent problem in MLR. One can exclude variables $K > M$ that are not significant; however, it is not a guaranteed solution [42].

In addition, where a situation like mixture design is considered, the **X**-variables could be component proportions, r_i, in the mixture that are not mutually independent ($i.e. \sum_{i=1}^{N} r_i = 1$, where $r_i \geq 0 \; \forall \; \mathrm{i} = 1,2,\ldots,N$). This situation is referred to as **X** being rank deficient, collinearity, zero determinant, singularity, and ill-conditioned. In such a situation, the inversion of ($\mathbf{X}^T\mathbf{X}$) matrix may not exist leading to the estimation of **β** using OLS with large variances. Here, a statistical multivariate method provides the right tool to extract systematic variables and remove collinearity in the data set as described in Section 2.6.

In molecular products, an important objective is to find a chemical product that exhibits certain desirable or specified behavior. Assuming the model in Eq. (2.18), and the parameter estimates $\hat{\boldsymbol{\beta}}$, a new *x*-variable can be predicted from a desired *y*-variable such that

$$\left(y_{1\times L}^{T}\right)_{des} = \left(x_{1\times K}^{T}\right)_{new} \cdot \hat{\boldsymbol{\beta}}_{K\times L}, \tag{2.20}$$

The model inversion of Eq. (2.20) gives

$$\left(x_{1\times K}^{T}\right)_{new} = \left(y_{1\times L}^{T}\right)_{des} \cdot \left(\hat{\boldsymbol{\beta}}_{L\times K}^{T} \cdot \hat{\boldsymbol{\beta}}_{K\times L}\right)^{-1} \hat{\boldsymbol{\beta}}_{L\times K}^{T} \tag{2.21}$$

where, $\left(\hat{\boldsymbol{\beta}}^{T} \cdot \hat{\boldsymbol{\beta}}\right)^{-1} \hat{\boldsymbol{\beta}}^{T}$ is the generalized inverse or pseudo-inverse of $\hat{\boldsymbol{\beta}}$

Eq. (2.20) does not contain any information about the covariance structure within the manipulated **X** variables. Consequently, the solution given by Eq. (2.21) will not respect those previous structural relationships when solving for the new conditions x_{new}^{T}. Therefore, the standard regression model (Eq. (2.20)) and its inversion (Eq. (2.21)) possess serious limitations in solving similar problems. Multivariate projection methods like PCA and PLS become indispensable tools in dealing with such difficulties and can be used for exploration, calibration, and classification of multivariate data.

2.6 Latent Variable Models

The most effective tools in multivariate data analysis are Principal Component Analysis (PCA) and Partial Least Square (PLS). PCA and PLS are decomposition techniques that compress a large quantity of data and extract the information by projecting them into a low-dimensional subspace that summarizes all the important information for analysis [43, 44, 45]. Then, further design work can be conducted in the reduced subspace.

If the data used is high dimensional and noisy, and the number of samples are small while developing a calibration model, there is the danger of over-fitting. In such cases, PCA (which can capture a dominant part of data variance) is more appropriate to reduce the data dimensionality and then train a regression model with the reduced latent variables. Principal component analysis is presented in this section as a variable reduction tool including important aspects such as pretreatment of data, validation, and

outlier detection that must be considered. Prior to PCA, data often needs to be pre-treated, in order to transform the data into a form suitable for analysis.

***Data Pre-processing**:* Pre-processing of data can make the difference between a useful model and no model at all. Pre-treating data is often employed to transform data into a form suitable for analysis. It is general practice to mean-center and scale the property variables prior to analysis [42, 46]. Variables often have substantially different numerical ranges. A variable with a large range has a large variance, whereas a variable with a small range has a small variance. Since PCA is a maximum variance projection method, it follows that a variable with a large variance is more likely to be expressed in the modeling than a low-variance variable [15].

In particular, the property descriptor data matrix $\mathbf{X}_{M\times K}$, consisting of M observations described by K descriptors, is mean-centered and *unit variance scaled (UV)* (also known as auto-scaled) across the M properties.

- ***Mean centering*:** The mean for each variable (each column) is calculated based on the entire sample and subtracted from each measurement (elements of matrix $\mathbf{X}_{M\times K}$).

$$x_{ij} = x_{ij} - \overline{x}_j \qquad \text{where, } \overline{x}_j = \frac{1}{M}\sum_{j=1}^{M} x_{i,j} \tag{2.22}$$

- ***Scaling*:** Variance scaling of the data to unity across the K descriptor properties can be accomplished by standardizing variables. For each measurement, this is done by dividing mean centered data by the standard deviation (s_j).

$$x_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j} \qquad \text{where, } s_j^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(x_{i,j} - \overline{x}_j\right)^2 \tag{2.23}$$

Auto-scaling puts all variables on equal footing; i.e., all variables have the same chance of entering the model and of taking part in the model. PCA performed on auto-scaled data is referred to as a correlation PCA.

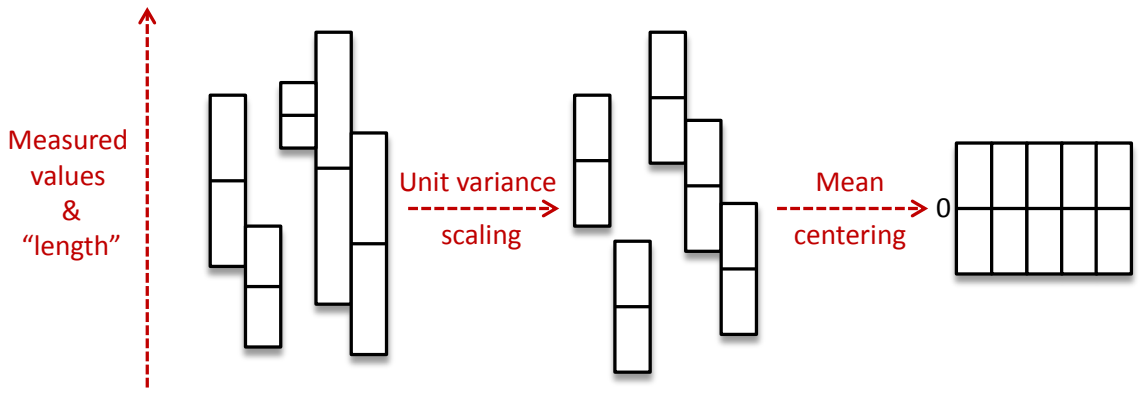


Figure 2.10: Unit variance scaled and mean-centered variables.

Figure 2.10 is a simple geometrical understanding of UV-scaling and mean-centering of variables [15]. In this representation, each bar corresponds to one variable and the short horizontal line inside each bar represents the mean value. The length of a bar (vector) is equal to its standard deviation (square root of variance). After UV-scaling, we get a shrinking of "long" variables and a stretching of "short" variables. By putting all variables on a comparable footing, no variable is allowed to dominate over another because of its length. Prior to any pre-processing, the variables have different variance and mean values. After UV-scaling, the "length" of each variable is identical; however, mean values still remain different. After mean-centering, mean values are zero (i.e., the centroid of the whole data set is zero). This improves the interpretability of the model developed.

However, it must be noted that in some cases, it is not necessarily advantageous to use auto-scaling, and some other choice might be more appropriate. Also, the **X**- and **Y**-variables can be scaled differently because the regression coefficients absorb the differences in scaling.

2.6.1 Principal Component Analysis (PCA)

Principal component analysis is a factor analysis method that is widely used in identification of systematic patterns in data and provides visualization of multivariate data by using as few variables as possible. It linearly maps multi-dimensional data onto lower dimensions with minimum loss of information [44]. The goals of PCA are to:

- extract the most important information from the data table,
- compress the size of the data set by keeping only this important information,
- simplify the description of the data set, and
- capture and analyze the structure of the variables.

In order to achieve these goals, PCA transforms a set of correlated variables into a new set of uncorrelated ones, known as principal components (PCs) such that

- the first PC is the linear combination of the standardized original variables that have the greatest possible variance,
- each subsequent PC is the linear combination of the standardized original variables that have the greatest possible variance, is orthogonal and has zero correlation with all previously defined PCs.

The orthogonality constraint imposed by the mathematics of PCA ensures that each variance-based axis is independent. PCs are arranged in order of decreasing eigenvalues. First PC is the most informative. Figure 2.11 shows the dimensionality reduction of original data to a low dimensional subspace using PCA which is much easier to visualize and analyze.

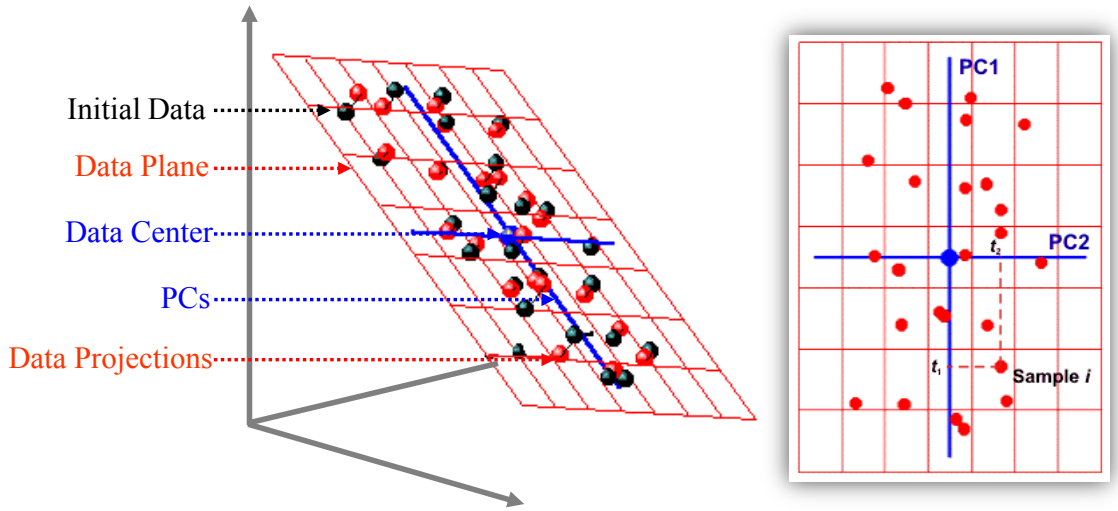


Figure 2.11: Projection of higher dimensional data onto a lower dimensional subspace.

Note: In the following discussion, matrices are denoted by capital bold characters (**X**, **Y**), column vectors by small italic characters (*t*), and row vectors by transpose vector (p^T).

Using PCA, a structural descriptor data set of molecular architecture information or a process condition matrix ($\mathbf{X}_{M\times K}$), representing M observations of K variables, can be decomposed to

$$\begin{aligned}\mathbf{X}_{M\times K} &= t_1 \cdot p_1^T + t_2 \cdot p_2^T + ...K \\ &= \sum_{i=1}^{K} t_i \cdot p_i^T \\ &= \mathbf{T}_{M\times K} \cdot \mathbf{P}_{K\times K}^T\end{aligned} \qquad \textbf{(2.24)}$$

where, **T** = the score matrix with mutually orthonormal columns

P = the loading matrix with mutually orthonormal columns

Principal components (PCs) are new lines that best approximate the data in the least squares sense and passes through the average point [15]. When two PCs are derived they, together, define a plane as seen in Figure 2.11). The score matrix (**T**) represents the projections of the data onto this

line in order to get a coordinate value along the PC-line. The new coordinate value is known as a score (t_i). The loadings define the orientation of the PC plane with respect to the original **X**-variable. The loading matrix (**L**) contains the coefficients in the linear combination of the original variable defining the *principal components* (*PCs*).

PCA-based soft models are both linear and additive. The loadings unravel the *magnitude* (large or small correlation) and the *manner* (positive or negative correlation) in which the measured variables contribute to the *scores*. Together the scores and loadings describe the principal components of the data set.

Normally, the first three A ($A \ll K$) PCs capture most of the variance in the data (80-90% of total variance [46]). By retaining only the first A PC's the **X** matrix can be approximated by:

$$\begin{aligned}\hat{\mathbf{X}}_{M\times K} &= t_1 \cdot p_1^T + t_2 \cdot p_2^T + ...A \\ &= \sum_{i=1}^{A} t_i \cdot p_i^T \\ &= \mathbf{T}_{M\times A} \cdot \mathbf{P}_{A\times K}^T\end{aligned} \qquad \textbf{(2.25)}$$

In order to characterize the properties of the observations one measures variables. Observations are comprised of two parts: signal and noise. Signal describes the property or effect of interest, and noise is everything else. Using methods based on variance such as PCA, multivariate data can be separated into signal and noise. Using Eq. (2.25), matrix **X** can be reconstructed as

$$\begin{aligned}\mathbf{X}_{M\times K} &= 1 \cdot \bar{x}^T + \hat{\mathbf{X}}_{M\times K} + \mathbf{E}_{M\times K} \\ &= 1 \cdot \bar{x}^T + \sum_{i=1}^{A} t_i \cdot p_i^T + \mathbf{E}_{M\times K} \\ &= 1 \cdot \bar{x}^T + \underbrace{\mathbf{T}_{M\times A} \cdot \mathbf{P}_{A\times K}^T}_{Structure} + \underbrace{\mathbf{E}_{M\times K}}_{Noise}\end{aligned} \qquad \textbf{(2.26)}$$

Here, $1 \cdot \bar{x}^T$ represents the variable average which originates from the pre-treatment step. This way, the data matrix $\mathbf{X}_{M\times K}$, containing K highly correlated variables, is transformed into the *score* matrix $\mathbf{T}_{M\times A}$, containing only A (where, $A \ll K$) mutually independent latent variables, which are linear combinations of the original K variables, have better properties (orthogonality) and also span the multidimensional space of $\mathbf{X}_{M\times K}$. The residual matrix (**E**) comprises the distances of the original variables to their projection onto the principal components.

Figure 2.12 graphically represents how scores and loadings form the **T·P** (structure) part of the PC model equation represented by Eq. (2.24).

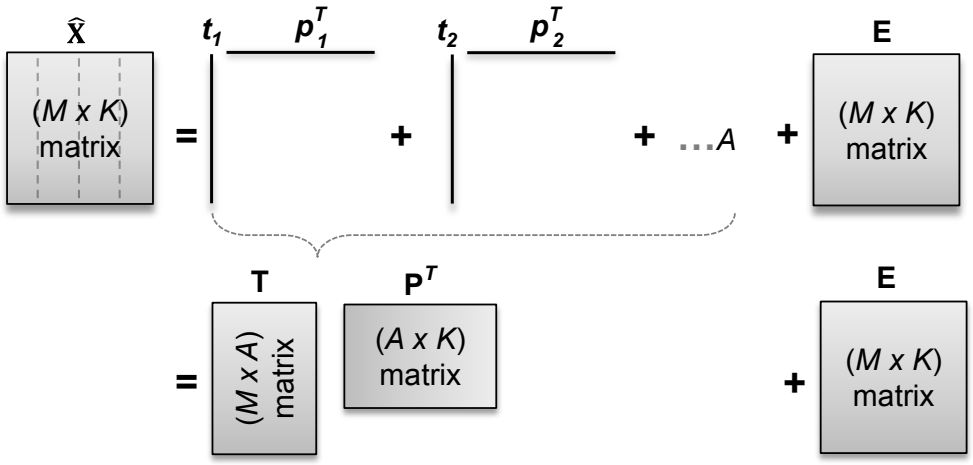


Figure 2.12: PCA decomposition of X matrix.

Here vectors are designated as column vectors and the corresponding transposed vectors are designated as row vectors. The dashed lines in the matrix indicate the mean centering and scaling direction.

2.6.1.1 Number of Principal Components

Before calibration of a model from PCs, it is important to determine the number of components necessary to extract the most relevant information from a data matrix. The errors calculated for the calibration set decrease

continuously as the number of components increase. Two methods [47] to help choose the number of components are to:

- plot the eigenvalues according to their size (the so called *scree plot*). The plot provides a visual aid for deciding at what point in this graph (often called an *elbow*) including additional components no longer increases the amount of variance accounted for by a nontrivial amount (slope of the graph goes from steep to flat). Keep only the components before the elbow. For example, the scree plot represented in Figure 2.13 suggests that using three PCs is appropriate.
- keep only the components whose eigenvalue is larger than the average eigenvalue. For correlation PCA, this means to keep only the eigenvalues larger than 1.

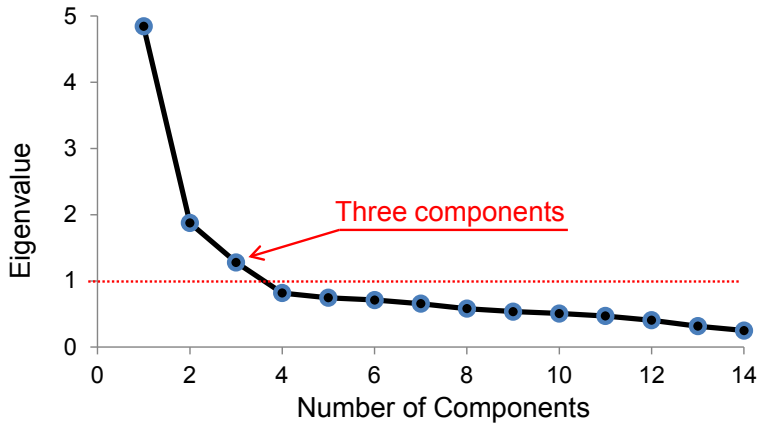


Figure 2.13: Scree plot of the correlation matrix.

One must be aware of the fact that unwanted variability in the data set, such as random noise, may also be taken into account by a model constructed with too many PCs. The model is said to be over-fitted, showing excellent results for evaluating samples belonging to the calibration set but failing on prediction of an external validation set.

2.6.1.2 Scores Plot

The scores plot shows correlations between observations, measurements or responses. Basically, it helps answer questions like:

- are observations related to each other?

  Responses close to each other have similar properties, whereas those far from each other are dissimilar with respect to descriptor variable profile (Figure 2.14 (a)).
- are there any groups or trends?

  In Figure 2.14 (a), data represented by different symbols (circle, triangle and square) represent a group of responses with similarity in descriptor properties.

2.6.1.3 Loading Plot

The loadings plot shows correlations between variables. It helps answer questions like:

- how the descriptor variables are correlated?

  Descriptor variables contributing similar information are grouped together, that is, they are correlated. For instance, in Figure 2.14 (b), variables 15 and 28 are correlated.
- which variables are influential?

  If the variables are in the same quadrant, they are positively correlated, whereas variables in opposite quadrants (opposite side of plot origin) are negatively (inversely) correlated. For instance in Figure 2.14 (b), when the value of one variable in upper-right hand corner (say variable 13) increases or decreases, the value of the other variable (say variable 16) has a tendency to change in the same way. Moreover, as the distance of a variable from the plot origin increases, the stronger is the impact of this variable on the model. This suggests that the variables 6, 13, 16, 44, 2, and 5 (from Figure 2.14 (b)) separate the three response groups (in Figure 2.14 (a)).

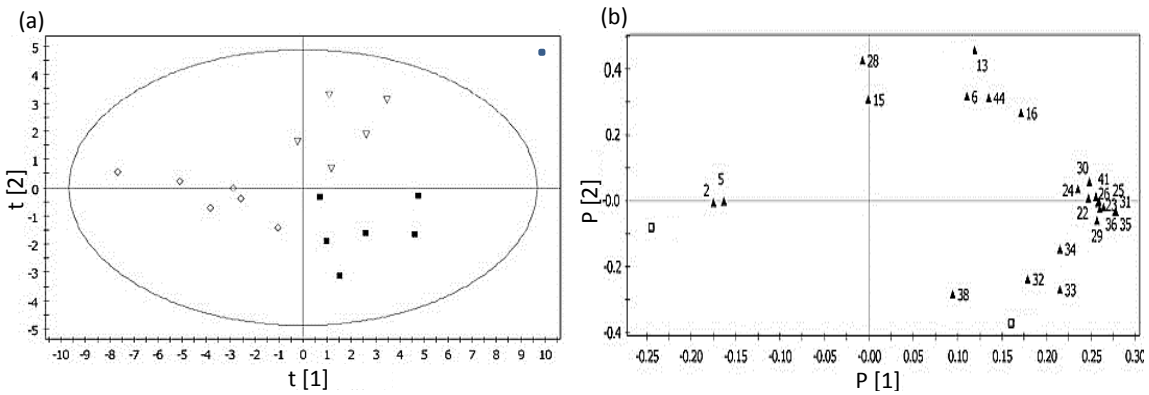


Figure 2.14: Principal component analysis: Score plot (a, left) of t_1/t_2 and loading plot (b, right) of p_1/p_2. The ellipse represents the Hotelling T^2 with 95% confidence in score plot.

Figure 2.14 depicts score and loading plots obtained from PCA. Comparing the loadings plot to the scores plot helps one understand how the variables relate to the observations.

2.6.1.4 Outlier Removal

Real data is rarely homogeneous (outlier-free) and therefore most statistical methods require removal of outliers prior to the calibration of a model. Outlying samples may have a huge influence on the calibration of a model and may decrease its predictive ability. Outlier identification is based on distance from data centroids. *Hotelling's* T^2 can be used to detect outliers inside the model space [48]. Hotelling's T^2 is a multivariate generalization of the univariate student's *t-test*, and provides a check for observations obeying to multivariate normality. When this statistic is used in conjunction with a score plot, Hotelling's T^2 defines a 95% or 99% tolerance region. In Figure 2.14 (a), one observation can be considered to be an outlier as it may not belong to the majority of the sample population and reveals a mistake in the property value obtained.

2.6.2 Principal Component Regression (PCR)

After applying PCA to the $\mathbf{X}_{M\times K}$ block for variable reduction, regression can be used to predict a particular quantitative characteristic (like attributes) as a function of score vectors. The multi-linear regression (MLR) relationship between the principal component scores $\mathbf{T}_{M\times A}$ and the attribute properties $\mathbf{Y}_{M\times L}$ can be developed using a PCR model as:

$$\hat{\mathbf{Y}}_{M\times L} = \mathbf{T}_{M\times A}\cdot\mathbf{B}_{A\times L}, \qquad \text{where, } \hat{\mathbf{B}} = \left(\mathbf{T}^T\cdot\mathbf{T}\right)^{-1}\cdot\mathbf{T}^T\cdot\mathbf{Y} \tag{2.27}$$

Unlike, in OLS where the columns of **Y** are regressed onto the large and highly correlated columns of **X**, in PCR the columns of **Y** are regressed onto the reduced and mutually independent latent variables **T**.

For any desired $(1\times L)$ vector of Y-variable, $(y^T)_{des}$, one can compute a $(1\times A)$ vector of new latent variable scores as

$$\left(t^T_{1\times A}\right)_{new} = \left(y^T_{1\times L}\right)_{new}\cdot\left(\hat{\mathbf{B}}^T_{L\times A}\cdot\hat{\mathbf{B}}_{A\times L}\right)^{-1}\hat{\mathbf{B}}^T_{L\times A} \tag{2.28}$$

and then predict the $(1\times K)$ vector of new **X**-variables as

$$\left(x^T_{1\times K}\right)_{new} = \left(t^T_{1\times A}\right)_{new}\cdot p_{A\times K} \tag{2.29}$$

Notice that Eq. (2.27) has the same structure as Eq. (2.18), however, instead of finding K variables to estimate $(x^T)_{new}$, now only A (where, $A \ll K$) latent variables has to be found to estimate $(t^T)_{new}$ thereby achieving reduction in dimension of the equation system involved. Also, since Eq. (2.25) preserves the covariance structure of **X,** the new **X**-variable found will be consistent with the past ones. More details can be found in [49].

2.6.3 Partial Least Squares (PLS)

PLS is a regression extension of principal component analysis (PCA). It generalizes and combines features from PCA and multiple linear regressions (MLR). Besides, relating two data matrices, **X** and **Y**, PLS also models the common structure between them thereby, giving richer results than the traditional multiple regression approach. PLS regression modeling has been described extensively in the open literature and can be found in Erikson *et al.* [15], Gabrielsson *et al.* [35], Geladi and Kowalski [42], among others. While this approach ensures the best possible correlation between the two data sets, it does not guarantee to best describe the **X** and **Y** data individually.

Figure 2.15 shows a PLS model being generated between the descriptor data matrix **X**, which could be molecular descriptors or property descriptors, and response data matrix **Y**, which could be attribute or property information. The method fits two "PCA-like" models at the same time, one for **X** and one for **Y**. However, the projections differ from those obtained with PCA on both blocks separately. The outer relation for **X** and **Y** block are:

$$\mathbf{X}_{M\times K} = 1\cdot\bar{x}^{T} + \mathbf{T}_{M\times A}\cdot\mathbf{P}_{A\times K}^{T} + \mathbf{E}_{M\times K} \qquad \textbf{(2.30)}$$

$$\mathbf{Y}_{M\times L} = 1\cdot\bar{y}^{T} + \mathbf{U}_{M\times A}\cdot\mathbf{V}_{A\times L}^{T} + \mathbf{F}_{M\times L} \qquad \textbf{(2.31)}$$

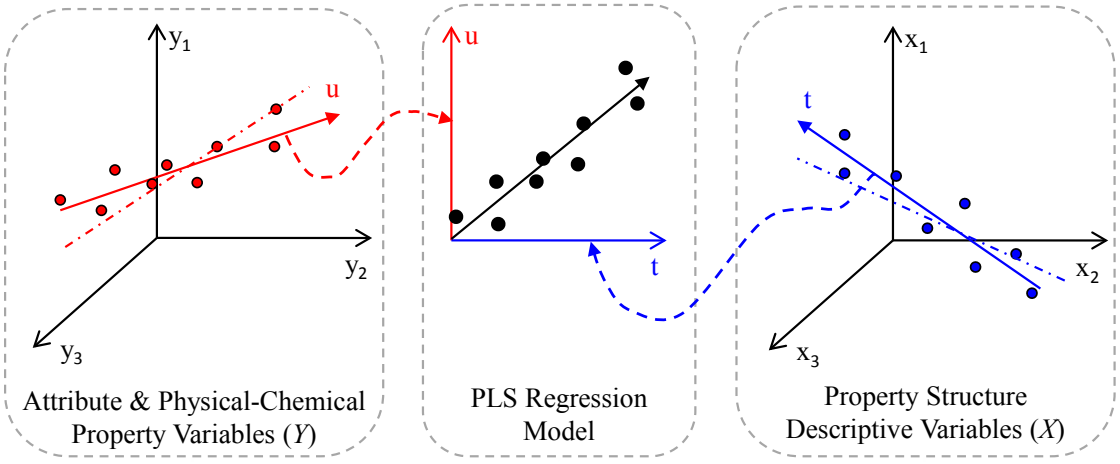


Figure 2.15: PLS regression on descriptive (*X*) and response (*Y*) variables.

The projections of **X** and **Y** are then connected through the inner relation

$$\mathbf{U}_{M\times A} = \mathbf{T}_{M\times A} + \mathbf{G}_{M\times A} \tag{2.32}$$

where, **G** is a residual and the regression coefficient is one.

The score plot in Figure 2.15 shows a linear relationship between predictors and the responses, however, non-linearities may exist. The PLS score plot *u*/*t* shows linear correlation structure between the predictors and the responses. The dash-dot line is the projection if PCA was performed on *X* and *Y* individually. In PLS, besides loadings **P** and **V**, there are additional loadings called weights $\mathbf{W}^*$ which express the correlation between **U** and **X** and are used to calculate **T** such that

$$\mathbf{T}_{M\times A} = \mathbf{X}_{M\times K}\mathbf{W}^*_{K\times A} \tag{2.33}$$

The prediction of **Y** can be obtained from the PLS model as:

$$\hat{\mathbf{Y}}_{M\times L} = \mathbf{T}_{M\times A}\cdot\mathbf{V}^T_{A\times L} = \mathbf{X}_{M\times K}\cdot\left(\mathbf{W}^*_{K\times A}\cdot\mathbf{V}^T_{A\times L}\right) = \mathbf{X}_{M\times K}\cdot\hat{\mathbf{B}}_{K\times L} \tag{2.34}$$

PCR and PLS can be considered standard calibration techniques for several spectroscopic techniques, among many. The main advantage of these techniques is to avoid collinearity problems thus allowing one to work with a number of variables that is greater than the number of samples. A comparison between these two techniques reveals similar results in terms of prediction ability in multivariate calibration, with no significant difference being reported when both employ the optimized number of principal components (PCs) [50]. PCR yields lower accuracy (degree of closeness of a measured value to the actual value) but higher precision (degree of closeness of the measured values to each other) than PLS. The basis of the model development in this dissertation will be based on PCR, however, PLS will be used for few qualitative applications.

2.6.4 Model Validation

It is important to validate the predictive ability of the developed model in addition to having a good fit. Validation can be distinguished between two types:

- ***Internal validation*:** where the calibration data is also used as validation data. This method is applicable when a proper validation sample is not available, or is not used. Often the calibration sample is separated into a learning set (for calibration) and a testing set (for validation) [15, 51]. A popular internal validation technique is the *leave-one-out* cross-validation technique in which one sample is used for validation and the remaining samples are used for calibration (one-at-a-time method). This is repeated for each sample. *Predicted residual sum of squares* (*PRESS*) is computed as:

$$PRESS = \sum_{i=1}^{K} \left(y_i - \hat{y}_i \right)^2 \qquad \textbf{(2.35)}$$

The number of components giving a minimum *PRESS* is the right number for the model that gives optimal prediction. The *root mean square error of cross-validation* is,

$$RMSECV = \sqrt{\frac{PRESS}{K}} \qquad \textbf{(2.36)}$$

The smaller the *PRESS* the better the quality of the estimation for a developed model. Details can be found in Montgomery *et al.* [51].

- ***External validation*:** is based on a new or independent set of validation data to evaluate the predictive ability of the previously developed model from the training data set. The external validation errors are often presented as the *root mean square error of prediction* (*RMSEP*), which is calculated as:

$$RMSEP = \left(\frac{\sum_{i=1}^{K} (y_i - \hat{y}_i)^2}{K} \right)^{1/2} \quad \textbf{(2.37)}$$

Including more latent variable terms in the final model will always increase the *goodness of fit*, R^2, regardless of whether the additional variable is statistically significant or not. However, a large value of R^2 does not necessarily imply that the regression model is a good one. Because R^2 always increases as we add terms to the model, *adjusted* R^2 statistics must also be considered. The R^2_{adj} is particularly useful in the selection stage of model building.

$$R^2_{adj} = 1 - \left(\frac{K-1}{A-1} \right) \left(1 - R^2 \right) \qquad \text{where, } R^2 = 1 - \frac{SS}{SS_{total}} \quad \textbf{(2.38)}$$

where, SS is the residual sum of square.

Unlike R^2, the R^2_{adj} increases only if the necessary terms improve the model more than could be expected by chance. Therefore, when R^2 and R^2_{adj} differ dramatically, there is a good chance that non-significant terms have been included in the model [51]. The R^2_{adj} can be negative, and will always be less than or equal to R^2.

An individual principal component (*PC*) generated by PCA or PLS is considered significant if its cross-validated Q^2 value is greater than zero. Q^2 is calculated according to

$$Q^2 = 1 - \frac{PRESS}{SS_{total}} \quad \textbf{(2.39)}$$

The overall significance of each PCA or PLS model is evaluated in terms of $Q^2(cum)$ as [15]:

$$Q^2(cum) = \left(1 - \prod \left(\frac{PRESS}{SS} \right) \right) \quad \textbf{(2.40)}$$

2.7 Principal Properties in Cluster Space

In order to solve a design problem in a single domain, all the physico-chemical attribute properties of interest have to be converted to principal properties by using the regression coefficients from the calibration model [27].

$$\mathbf{T}_{M\times A} = \mathbf{Y}_{M\times L}\cdot\mathbf{B}^{-1}_{L\times A} \qquad \text{where, } \mathbf{B}^{-1}_{L\times A} = \begin{cases} \left(\hat{\mathbf{B}}^{T}_{L\times A}\cdot\hat{\mathbf{B}}_{A\times L}\right)^{-1}\hat{\mathbf{B}}^{T}_{L\times A}, & \text{if } L > A \\ \hat{\mathbf{B}}^{T}_{L\times A}\left(\hat{\mathbf{B}}_{A\times L}\cdot\hat{\mathbf{B}}^{T}_{L\times A}\right)^{-1}, & \text{if } L < A \end{cases} \tag{2.41}$$

$\mathbf{B}^{-1}$ is a generalized inverse or pseudo-inverse of a matrix $\hat{\mathbf{B}}$ [52].

In order to utilize the latent variables (LVs) in the property clustering algorithm, it is important that the LV structures follow a linear mixing rule. This can be achieved by standardizing the data structure to obtain a new matrix, $\mathbf{Q}_{M\times A}$. Rearranging the data decomposition represented by Eq. (2.25), we get:

$$\mathbf{T}_{M\times A} = \mathbf{X}_{M\times K}\cdot\mathbf{P}_{K\times A} \tag{2.42}$$

If the loadings $\mathbf{P}_{KxA}$ are thought of as the pure values of the principal components, then scores $\mathbf{T}_{MxA}$ serves as the predicted mixture properties represented by $\mathbf{\Psi}_{Mix}$ in Eq. (2.1). Here, the mixture fractions of the multivariate descriptor data $\mathbf{X}_{MxK}$ must sum to one across descriptor variables *K*. To achieve this form, the latent variable structure is standardized by dividing $\mathbf{T}_{MxA}$ and $\mathbf{X}_{MxK}$ in Eq. (2.42) by the sum total of the property descriptors of each experimental run $\mathbf{S}_{KxM}$ such that:

$$\mathbf{Q}_{M\times A} = \mathbf{U}_{M\times K}\cdot\mathbf{P}_{K\times A} \tag{2.43}$$

Where, $\mathbf{Q}_{MxA} = \mathbf{T}_{MxA}/S_M$, $\mathbf{U}_{MxK} = \mathbf{X}_{MxK}/S_M$, and $S_M = \sum_{i}^{H} x_i$.

A powerful chemical product design framework is achieved, by integrating latent variable methods within property cluster domain and by

formulating two reverse problems. The first reverse problem identifies product quality/performance requirements and second identifies mixing conditions (involving selection of materials, blend ratios, or process operation conditions) when considering mixture design, and substructural molecular building blocks when considering molecular design. The following chapters will demonstrate the mixture and molecular design solution framework developed by combining methods and tools presented in this chapter.

2.8 Computer-Aided Design using QSPR and cGCM

This section describes how the tools and techniques presented in this chapter are combined within the computer-aided approach to facilitate investigation of chemical product formulations prior to experimentation and simulation. When appropriate property models are available to describe and predict the target product properties, computer-aided molecular design methodologies are utilized to solve the design problems systematically and efficiently. The computer-aided molecular design problems require: (1) prediction of properties from molecular structure (solution to forward problem) and (2) identification of optimized molecular structures that meet given a set of property values (solution to reverse problem) [1, 9, 11, 24- 27, 53]. Figure 2.16 shows the two required solutions of CAMD.

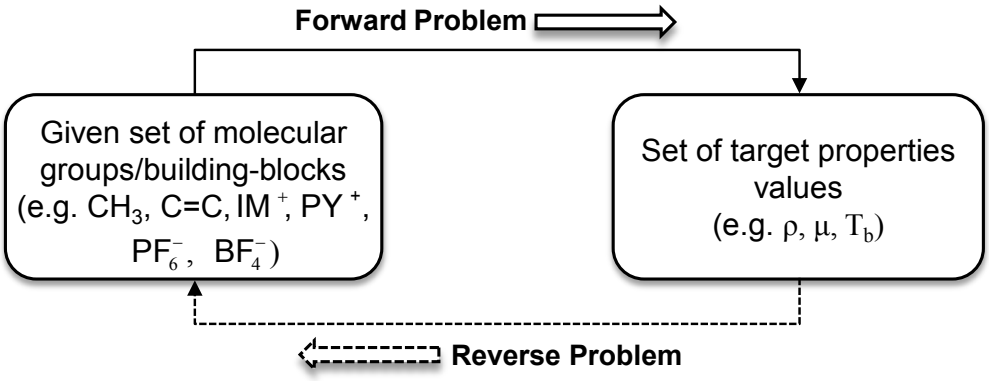


Figure 2.16: Forward and reverse problems in computer-aided molecular design.

2.8.1 Solution to Reverse Problem

The molecular design problem presented in this dissertation follows a reverse problem of property prediction from molecular structures. In this approach, it is required to enumerate structured molecules from a given set of representative molecular building blocks or fragments corresponding to target physical-chemical properties estimated by property models. First, the values of the target properties of desired molecule needed for a specific application are determined *a priori*. Lower and upper bounds on each property, y_j, of molecule, i, are obtained either from process design problem (see Section 2.2.1) or from product property requirements/constraints.

$$y_j^{\min} \le y_{ij} \le y_j^{\max} \tag{2.44}$$

Mapping the design problem formulation from property to cluster space (described in Section 2.2.2) enables visualization of a problem and its solution when three properties are concerned. The normalized property operator of molecule i using corresponding reference values give:

$$\Omega_j^{\min} \le \Omega_{ij} \le \Omega_j^{\max} \tag{2.45}$$

where Ω_j is defined by Eq. (2.4).

Second, brute-force search or exhaustive searches are performed using numerous permutations or combination of atoms and molecular groups, to systematically enumerate all possible molecular structures, and mixtures of molecules, that satisfy specifications in terms of the normalized target property operator values in Eq. (2.45). The maximum number of similar groups N_g^* are predefined to limit the size of a generated molecule. The possible number n_g^* of similar molecular fragments/groups of type g that can be included in the design can be estimated by taking the minimum of the nearest integer value of $\left(\Omega_j^{max}/\Omega_j\right)$ and N_g^*.

$$n_g^* = \min\left(\frac{\Omega_j^{\max}}{\Omega_j} : N_g^*\right) \quad \textbf{(2.46)}$$

where Ω_j^{max} is the target maximum latent property value of property j.

It must be noted that exhaustive generate-and-test algorithms are adequate to solve problems with small size. For problems with slightly larger search space, reduction in the search space (reducing the set of candidate solutions to a manageable size) using problem-specific heuristics can make the algorithm more efficient [54, 55]. However, for large scale molecular design problems, stochastic or evolutionary search algorithms such as genetic algorithm (GA), are preferred [9].

2.8.2 Solution to Forward Problem

In order to achieve predictive property models, quantitative structure-property relations (QSPR) are utilized where the training set molecular descriptors are generated in terms of infrared (IR) frequencies. Since molecular descriptors are obtained through characterization techniques based on IR spectroscopy (described in Section 2.4.1), the group based property estimation method is termed *characterization-based group contribution method* (cGCM) [27]. Since the combination of molecular groups can be done in an infinite number of ways, the search space and the size range of enumerated molecules can be reduced by constraining the number of occurrence of similar and dissimilar types of groups within minimum and maximum limit.

$$0 \le n_g^* \le N_g^* \quad \textbf{(2.47)}$$

$$2 \le n_g \le N_g \quad \textbf{(2.48)}$$

Note that a minimum number of two dissimilar groups, n_g, must be selected to form a structurally feasible molecule.

When order of selection does not matter and repetition of similar groups is allowed, the total number of possible candidate molecules that can be generated by selecting n_g groups from a set of N_g groups is given by:

$$\text{Candidate}_{\text{total}} = \sum_{n_g=2}^{N_g} {}^{N_g}C_{n_g} = \sum_{n_g=2}^{N_g} \frac{\left(N_g + n_g - 1\right)!}{n_g!\left(N_g - 1\right)!} \tag{2.49}$$

As the number of groups gets high, the problem of combinatorial explosion arises. In order to reduce the combinatorial problem, atoms and several first order groups (except some terminal groups like CH_3) are clustered into *meta-groups*. In group-based property estimation methods, these meta-groups are treated as first order groups. Such groups capture inter- and intra-atomic and group interactions. The property operators of molecules using first order meta-groups-based cGCM are estimated as:

$$\Omega_j^{Mix} = \sum_{g=1}^{N_g} n_g \cdot \Omega_{jg} \tag{2.50}$$

where, Ω_j^{Mix} = the mixture property operator values of property j

n_g = the number of occurrences of dissimilar molecular fragment g

N_g = the maximum number of possible appearance of dissimilar fragments (user defined), and

Ω_g = the property contribution of the appeared group in the formulated molecule.

When the predicted property values using Eq. (2.50) of enumerated molecules from the combinatorial building blocks satisfy the property constraints in Eq. (2.45), the set of molecules are considered and selected as a candidate solution. Then the enumerated candidate molecules are screened for structural constraints to ensure that a stable and connected molecule was formed. One such structural constraint is to check the number of unused bonds in a generated molecule, i.e. the free bond number (*FBN*) [24].

$$FBN = \left(\sum_{g=1}^{N_g} n_g \cdot FBN_g \right) - 2 \cdot \left(\sum_{g=1}^{N_g} n_g - 1 \right) - 2 \cdot N_{Rings} \quad \textbf{(2.51)}$$

where, FBN_g = the unique free bond number associated with group g

N_{Ring} = the number of rings in the formulation.

For an electronically complete molecular formulation the free bond number must be equal to zero. A *FBN* of zero indicates that the electron valency shells of all atoms in the molecule have been satisfied, which, in most cases, indicates one of the minimum energy connectivity configurations of the atoms in the molecule.

Finally the feasible solution can be verified through more rigorous experimentation and/or computational studies based on molecular dynamics and quantum chemical calculations. The methodology presented in this dissertation provides a computationally efficient screening procedure to narrow down potential compounds with desired attributes from a large chemical space. Figure 2.17 is a schematic diagram that summarizes the steps involved with variable transformation, latent variable model calibration, and property integration to solve the computer-aided reverse design of chemical products in reduced space.

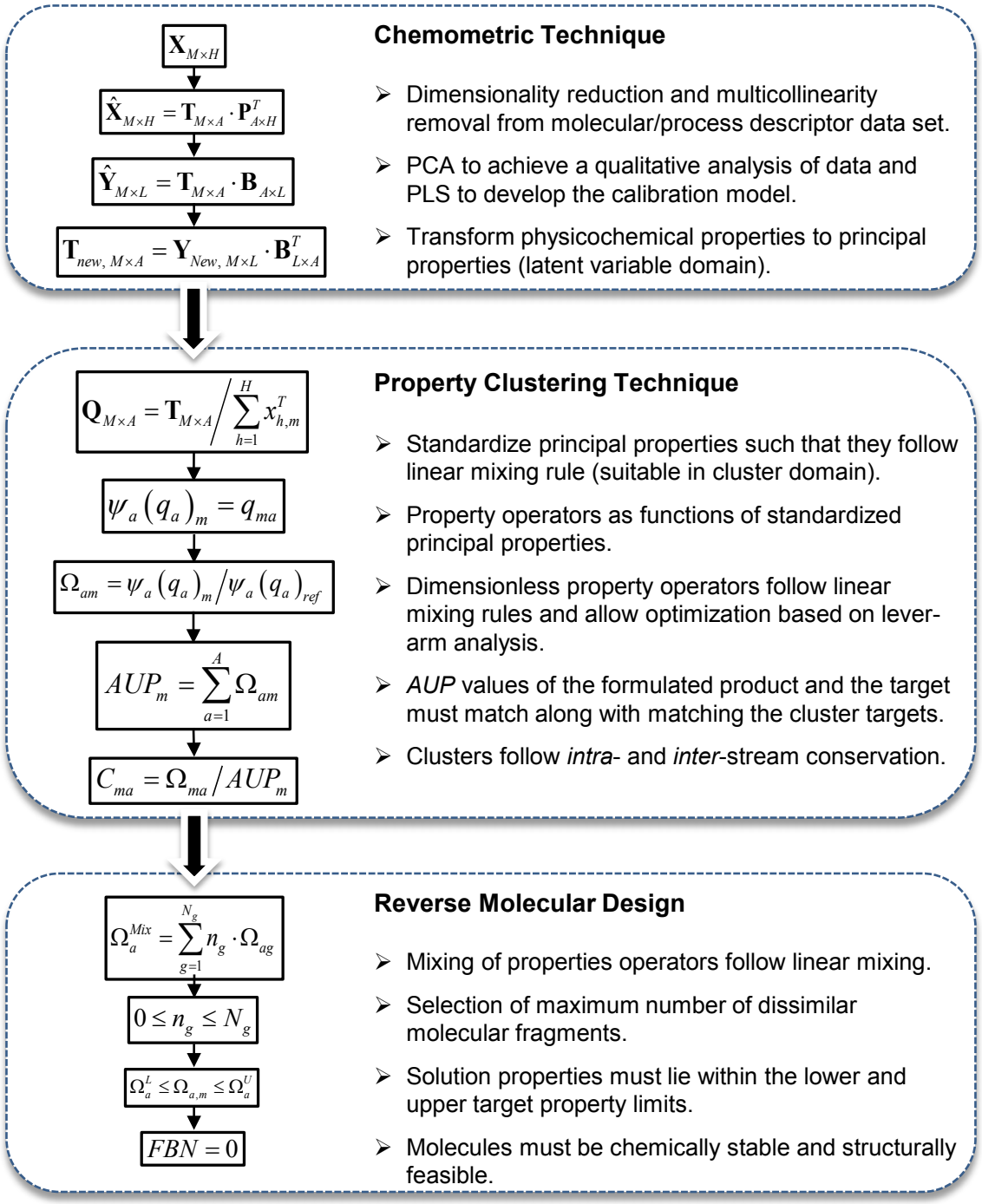


Figure 2.17: An overview on the methodology of multivariate characterization, modeling, and design.

CHAPTER 3

MIXTURE DESIGN

3.1 Introduction

Mixing/blending processes play an important role in today's manufacturing of value-added chemical products such as commodity chemicals, food, cosmetics, oil and pharmaceuticals. The product formulation design problem in industrial research and development is one such area where the analysis of mixture data could be utilized more effectively.

Mathematical models are commonly used to characterize a system, which is to be controlled or optimized by a set of variables, to study the effects of various factors and to make predictions about behavior. Generally, models can be classified into mechanism-driven models and data-driven models. Traditionally, the chemical engineering discipline has focused on mechanistic models that describe underlying phenomena with a system of differential and algebraic equations (DAEs). However, real-life industrial process systems, which are often complex and nonlinear with incomplete and/or uncertain data, cannot be adequately described by such models [18].

Data-driven models offer an alternative solution as we move from limited data, which used to be obtained through time-consuming experiments and simulations, to massive amounts of data that can be generated from analytical instruments, images, spectra, etc. (i.e. "data poor" to a "data rich" paradigm shift due to rapid instrumentalization of science and technology). This has been the trend in recent decades [18, 19, 35, 27, 41, 56, 57].

Industrial reality also suggests that a good theoretical process model is often not available. Before new experiments are conducted, historical process

data that encompass a wide spectrum of operating conditions and existing product grades can be utilized to achieve new and improved products (20). Today, tremendous amounts of diverse data are readily available from extensive monitoring of equipment, processes, and products at all scales (18). However, managing such abundance of complex data to build appropriate models for a specific application and exploring their effects on the final product properties remain major challenges.

3.2 Traditional Approach

In classical non-mixture designs such as factorial and response surface designs, all the factors are orthogonal or independent. This means that it is possible to freely choose the level of a factor regardless of the other factors' levels [58]. In general, the model parameters can be used to judge the effects of the mixture components. For example, consider a mixture experiment in which three components x_1, x_2, and x_3, were blended to form a product with response variable y. If $\hat{y} = 11.7x_1 + 9.4x_2 + 16.4x_3 + 19.0x_1x_2 + 11.4x_1x_3 - 9.6x_2x_3$ can adequately represent the response, it can be concluded that component 3 has the largest contribution to the highest response value because the coefficients are in the order $\hat{\beta}_1 > \hat{\beta}_2 > \hat{\beta}_3$. Furthermore, since $\hat{\beta}_{12}$ and $\hat{\beta}_{13}$ are positive, presence of components 1 and 2 or components 1 or 3 enhances the response. This is an example of *synergistic* blending effects. Components 2 and 3 have *antagonistic* blending effects because $\hat{\beta}_{23}$ is negative and the presence of both components works against the response of interest.

However, this freedom does not exist for mixture designs, because each component in a mixture is dependent upon the settings of the other component settings. In mixture design, the factors are interdependent and the effects of the factors on the responses are not separable. For example, for a mixture containing N components, if the proportions of the first N-1 components are defined, the proportion of the N^{th} component cannot be freely

chosen. If x_i is the percentage content of component i, the sum of all the mixture components is given by the fundamental mixture constraint relationship

$$\sum_{i=1}^{N} x_i = x_1 + x_2 + \cdots + x_N = 1.0 \qquad \textbf{(3.1)}$$

$$0 \le x_i \le 1.0 \qquad \forall\ i = 1, 2, ..., N \qquad \textbf{(3.2)}$$

Analyzing mixture data with multiple regressions necessitates a special model form to eliminate the mixture constraint (Eq. (3.1)). The Scheffe canonical models [59] and Cox polynomial models [60] are the two most commonly used to analyze mixture data with multiple regressions. Scheffe introduced canonical models of various orders by eliminating some terms from the complete polynomial model. Cornell [58] provides multiple standard references to regression, modeling and analysis of mixture data. A short review on Scheffe and Cox models is also included in Appendix B.

In most practical applications, mixture data can be noisy and highly collinear because of process or operational constraints (20). Although the Scheffe canonical models and the Cox polynomial models (a reparameterized and constrained version of the Scheffe model) eliminated the *true collinearity*, and enabled the use of multiple regressions for the estimation, the problem of *near collinearities* with mixture data remains. *Design of experiments* (DOE) with *response surface methods* is usually used to determine the optimum combination of chemical constituents that give a desired response using a minimum number of experimental runs [35]. While such a design approach is adequate for most experimental designs, it suffers from combinatorial explosion and visualization difficulties when dealing with multi-component mixtures [22,61]. Solvason *et al.* [61] presented a solution to the above problems by integrating the property clustering framework with existing mixture design techniques.

In addition, traditional mixture models are usually employed to investigate the relationships between the blend ratio matrix (**R)** and the final blend product property matrix (**Y)** only (Eq. (3.3)), given that the properties of the pure raw materials (**X**) and the process conditions used to manufacture them (**Z**) are already chosen. The preceding topics represent this situation. However, for the development of new products that meet target properties with minimum experiments and minimum total material cost, it is important to simultaneously take into account all three degrees of freedom (**X**, **R**, and **Z**) available in blending operations [57].

In general, mixture design approaches tend to treat the three problems as separate steps. A set of raw materials is selected usually based on the experimenters' best guess, and then a set of blending experiments with certain constant process conditions are run to see if the target properties can be achieved. If the results are not acceptable, another set of raw materials and/or process conditions are selected and the process repeated. Figure 3.1 depicts the traditional mixture design approach which may lead to many blending experiments, a very inadequate investigation of the large number of raw materials, and a very long development time.

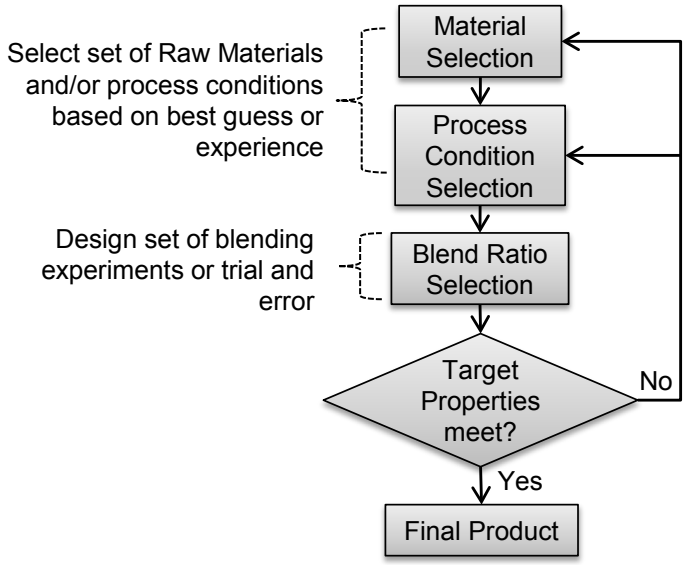


Figure 3.1: Traditional approaches in mixture design.

Statistical multivariate models such as principal component analysis (PCA), partial least squares (PLS), and neural networks (NN) provide powerful tools to extract systematic variables and remove both of the above collinearities in the data set, thus allowing one to work with a number of variables that is greater than the number of samples [15, 49, 56, 57].

Kettaneh-Wold [56] proposed the use of a partial least square (PLS) model for mixture data, and showed how it effectively deals with both the above collinearity problems, and can simultaneously incorporate process conditions. Recently, Muteki *et al.* [57, 62] added the relationship between the raw material properties and final blend properties that allows investigation of the effect of raw material properties on the final mixture product properties. However, it is important that the raw material properties selected are sufficiently correlated with the final blend product properties. The multivariate statistical methods presented in this chapter are based on Muteki's development on mixture-property models combined with the property clustering technique and reverse problem formulation. The nomenclatures are kept consistent from that in Muteki's papers for less confusion. In this chapter we shall call the **X**-variables *factors* or *predictors* and the **Y**-variables *qualities* or *responses*.

3.3 Multi-Block Data Structure

The data structure generally available on raw material property data and blending data in mixture design is shown in Figure 3.2 [57]. The raw material properties matrix, $\mathbf{X}_{N\times K}$, consists of N available raw materials with K number of properties. The blend ratio matrix, $\mathbf{R}_{M\times N}$, consists of M number of blends of N materials used in the formulation of the blends such that $\sum_{i=1}^{N} r_i = 1.0$, where $r_i \geq 0 \; \forall \; \mathrm{i} = 1,2,\ldots,N$. The process conditions matrix, $\mathbf{Z}_{M\times J}$, consists of J process conditions. The final response or property matrix, $\mathbf{Y}_{M\times L}$, consists of L properties measured on the final product. It must be noted that

there is no common dimension over the entire data matrix. However, $\mathbf{X}^{T}$ has an indirect relationship to **Y** through **R** as it has one dimension in common with **R** but no dimension in common with **Y**. It is referred to as *L*-or *T-shaped* data structure [57].

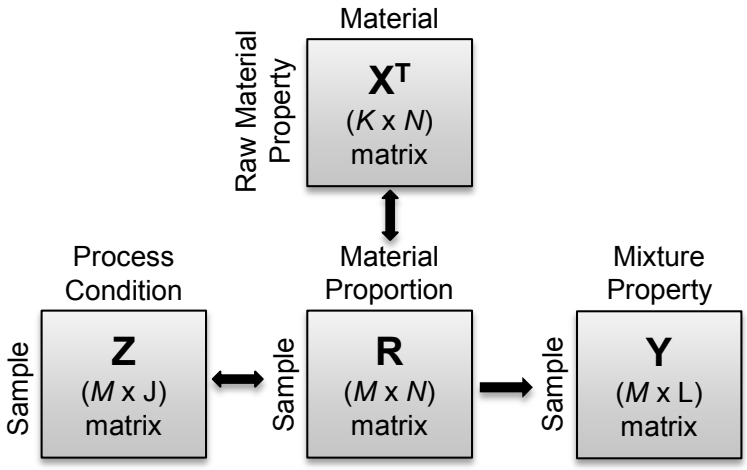


Figure 3.2: Data structure for three manipulative variable matrixes and a quality/response variable matrix.

3.4 Multi-Block Regression Models

Traditional mixture models such as the Scheffe and the Cox models [58], have commonly been used to model the relationship between **R** and **Y** as:

$$\mathbf{Y} = f(\mathbf{R}) + \varepsilon \tag{3.3}$$

Muteki *et al.* [57] used ideal mixing rules for combining the raw material properties matrix (**X)** and the blend ratios matrix (**R**) in order to relate all (**Z**, **R**, $\mathbf{X}^{T}$) to the **Y** matrix with a common dimension *M*.

$$\mathbf{Y} = f(\mathbf{X}_{mix}, \mathbf{Z}) + \varepsilon \qquad \text{where, } \mathbf{X}_{mix,M\times K} = \mathbf{R}_{M\times N} \cdot \mathbf{X}_{N\times K} \tag{3.4}$$

$\mathbf{X}_{mix}$ is the mixture-raw material mixture-property matrix. Muteki [57] demonstrated that the mixture-property models account for the similarities among the raw material properties and their effect on the blends. This can

successfully capture more inherent latent relationships between the mixture properties ($\mathbf{X}_{mix} = \mathbf{R} \cdot \mathbf{X}$) and the final product properties **Y** than those which exist simply between the ratios (**R**) and final blend properties (**Y**). When a property model does not follow linear mixing, then the model can be moved from its non-linear domain to a linear domain by a suitable transformation of the model formulation [34]. If the process operating conditions (**Z**) change between blending experiments, then the effect of these changes are easily accounted for by incorporating **Z** into the PLS models:

$$\mathbf{Y} = f(\mathbf{X}_{mix\ all}, \mathbf{Z}) + \varepsilon \qquad \text{where, } \mathbf{X}_{mix\ all} = \left[\mathbf{X}_{mix}\mathbf{Z}\right] = \left[\left(\mathbf{R} \cdot \mathbf{X}\right)\mathbf{Z}\right] \tag{3.5}$$

$\mathbf{X}_{mix\ all}$ is a matrix combining $\mathbf{X}_{mix}$ and **Z** in parallel. The above data combination can be better understood graphically in Figure 3.3. (The dotted lines in each block indicate the mean centering and scaling direction.)

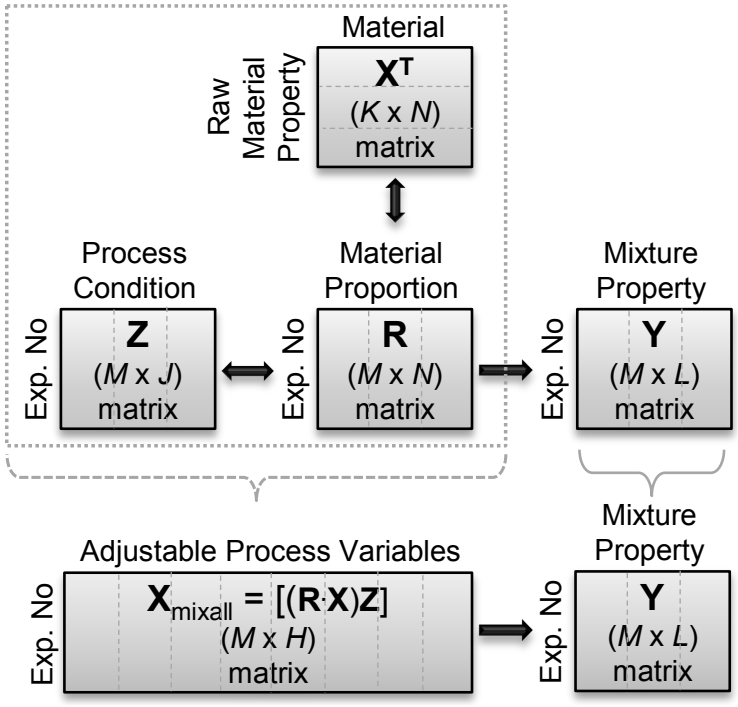


Figure 3.3: Data structure for combined manipulative variable matrixes and a quality/response variable matrix.

This simplifies the analysis and design by not having to differentiate between mixture and process variables and not having to assume independence of the factors when multivariate analysis techniques are used [56]. The multi-block PCR (Section 2.6.2) or PLS (Section 2.6.3) model can be used to obtain the relationship in Eq. (3.5). PLS regression is performed by projecting the $\mathbf{X}_{mix\ all}$ data and $\mathbf{Y}$ data onto a lower dimensional subspace:

$$\mathbf{X}_{mix\ all,M\times H} = \sum_{i=1}^{A} t_i \cdot p_i^T + \mathbf{E}_{M\times H} = \mathbf{T}_{M\times A} \cdot \mathbf{P}_{A\times H}^T + \mathbf{E}_{M\times H} \qquad \text{where, } \mathbf{T} = \mathbf{X}_{mix\ all} \cdot \mathbf{W}^* \tag{3.6}$$

$$\mathbf{Y}_{M\times L} = \sum_{i=1}^{A} t_i \cdot v_i^T + \mathbf{F}_{M\times L} = \mathbf{T}_{M\times A} \cdot \mathbf{V}_{A\times L}^T + \mathbf{F}_{M\times L} \tag{3.7}$$

The prediction of $\mathbf{Y}$ can be obtained from the PLS model as:

$$\hat{\mathbf{Y}}_{M\times L} = \mathbf{T}_{M\times A} \cdot \mathbf{V}_{A\times L}^T = \mathbf{X}_{mix\ all,M\times H} \cdot \mathbf{W}_{H\times A} \cdot \mathbf{V}_{A\times L}^T \approx \mathbf{X}_{mix\ all,M\times H} \cdot \hat{\mathbf{B}}_{H\times L} \tag{3.8}$$

The prediction of $\mathbf{Y}$ can be obtained from the PCR model as:

$$\hat{\mathbf{Y}}_{M\times L} = \mathbf{T}_{M\times A} \cdot \hat{\mathbf{B}}_{A\times L} \tag{3.9}$$

This way, the data matrix $\mathbf{X}_{mix\ all,M\times H}$, containing $H\ (= K + J)$ highly correlated manipulated (or predictive) variables is transformed into the score matrix, $\mathbf{T}_{M\times A}$, containing only A (where, $A < K$) independent latent variables, which are linear combinations of the original manipulated variables. The weights of this linear combination are captured in the loading matrices, $\mathbf{W}_{K\times A}^*$, $\mathbf{P}_{H\times A}$ and $\mathbf{V}_{L\times A}$. If new raw material properties, $(x^T)_{new}$, have to be predicted from the desired product quality specifications, $(y^T)_{des}$, then the inversion of the latent variable model gives:

$$\hat{x}_{new,1\times K}^T = \hat{y}_{des,1\times K}^T \cdot (\mathbf{B}^T \cdot \mathbf{B})^{-1} \cdot \hat{\mathbf{B}}_{L\times A}^T \cdot \mathbf{P}_{A\times K}^T \tag{3.10}$$

3.5 Proof of Concept Example – Starch Blending

The development of thermo-plastics from the mixing of starches, lactic acids and additives using latent variables in cluster space is presented as a case study to illustrate the method and concept described in this chapter. Data for the polymer blend problem was obtained from Muteki [63] and involves study of the influence of raw material properties ($\mathbf{X}^{\boldsymbol{T}}$), blend ratios (**R**) and process conditions (**Z**) on the product property matrix **Y**. This work employs multivariate data analysis techniques and the mixture-property model introduced by Muteki et al. (2006) then formulates and solves the mixture design problem in the cluster space.

3.5.1 Structure of starch blending data

The raw material data matrix (**X**) consists of 5 properties (Amylose content and 4 properties related to molecular weight distribution) on 3 starches. The blend matrix (**R**) consists of different blend ratios of 3 starches and 3 other materials (1 polylactic acid and 2 additives) in each blend of 28 mixtures. The process condition matrix (**Z**) consists of the molding temperature as the process variable. The final product property matrix (**Y**) consists of 4 polymer properties (tensile strength (*TS*), tensile modulus (*TM*), elongation at break (*EB*) and density (*Rho*)) for 28 mixture blends.

Figure 3.4 represents a more complex extension of the data structure for the **X** and **R** matrices shown in Figure 3.2. The raw material property data matrices result in a staircase type of the data structure and are not overlapped with each other because their raw material properties contain measurements on different variables due to the different classes of materials. These blend ratio matrices were designed by a D-optimal design, and therefore the conditioning is relatively good. In this case, only one class of raw materials has the property data information ($\mathbf{X}^{\boldsymbol{T}}_{starch}$), and the property data matrix ($\mathbf{X}^{\boldsymbol{T}}_{others}$) of the other materials (PLA and additives) is not

available [63]. The data structure of the **X** matrix for this problem of the mixture-property model using ideal mixing rule becomes $\mathbf{X}_{mix\,all} = [(\mathbf{R} \cdot \mathbf{X}_{mix})\mathbf{Z}] = [(\mathbf{R}_{starch} \cdot \mathbf{X}_{starch})\mathbf{R}_{other}\mathbf{Z}]$. It is also known that many polymer properties, such as the average molecular weight approximately follow ideal mixing rules [64].

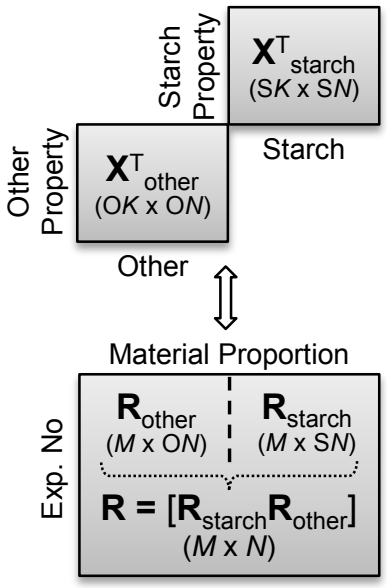


Figure 3.4: Data structure of X and R matrix for two classes of raw materials and their blend ratios.

Table 3.1 and Table 3.2 contain the data for the raw material properties and Table 3.3 contains mixture conditions and product properties [63].

Table 3.1: Starch material property data matrix.

Raw Material Property Data, X_{starch}					
Starch I.D.	**Amylose content,**	**Average molecular weights (×10^{-3} g/mol)**			
	wt%	***Mn***	***Mw***	***Mz***	***Mw/Mn (PDI)***
Starch1	70	41	214.5	463.7	5.23
Starch2	23	62.3	531.4	821.2	8.53
Starch3	0	90.1	722.2	1040	8.02

Table 3.2: Blending ratio and process condition data matrix.

	Material Blend Ratio						Process Condition
Mixture	R_{starch}			R_{other}			Z
I.D.	Starch 1	Starch 2	Starch 3	PLA	Additive 1	Additive 2	Molding temperature
1	0.50	0.00	0.00	0.00	0.50	0.00	150
2	0.41	0.00	0.00	0.18	0.41	0.00	150
3	0.41	0.00	0.00	0.18	0.41	0.00	170
4	0.33	0.00	0.00	0.33	0.33	0.00	150
5	0.33	0.00	0.00	0.33	0.33	0.00	170
6	0.23	0.00	0.00	0.54	0.23	0.00	150
7	0.23	0.00	0.00	0.54	0.23	0.00	160
8	0.23	0.00	0.00	0.54	0.23	0.00	170
9	0.39	0.00	0.00	0.17	0.39	0.05	150
10	0.32	0.00	0.00	0.32	0.32	0.05	150
11	0.22	0.00	0.00	0.51	0.22	0.05	150
12	0.16	0.00	0.00	0.63	0.16	0.05	150
13	0.36	0.00	0.00	0.36	0.24	0.05	150
14	0.36	0.00	0.00	0.36	0.24	0.05	170
15	0.38	0.00	0.00	0.38	0.19	0.05	150
16	0.38	0.00	0.00	0.38	0.19	0.05	170
17	0.36	0.00	0.00	0.36	0.18	0.09	150
18	0.36	0.00	0.00	0.36	0.18	0.09	170
19	0.00	0.52	0.00	0.22	0.26	0.00	150
20	0.00	0.52	0.00	0.22	0.26	0.00	170
21	0.00	0.40	0.00	0.40	0.20	0.00	150
22	0.00	0.40	0.00	0.40	0.20	0.00	170
23	0.00	0.26	0.00	0.61	0.13	0.00	150
24	0.00	0.26	0.00	0.61	0.13	0.00	170
25	0.40	0.00	0.00	0.40	0.20	0.00	150
26	0.40	0.00	0.00	0.40	0.20	0.00	170
27	0.00	0.00	0.40	0.40	0.20	0.00	150
28	0.00	0.00	0.40	0.40	0.20	0.00	170

Table 3.3: Mixture product quality data matrix.

	Product Quality, Y			
Mixture I.D.	**Tensile strength Mpa**	**Tensile modulus Mpa**	**Elongation at break %**	**Density g/cm³**
1	0.888	18.59	10.1	1.367
2	3.72	162.3	10.1	1.341
3	3.25	166.1	7.6	1.332
4	9.7	629.1	6.59	1.311
5	7.55	489.1	4.02	1.316
6	22.8	1522	7.59	1.296
7	23.8	1469	8.00	1.298
8	20.0	1430	4.21	1.294
9	3.03	134.5	11.00	1.345
10	8.73	568.9	7.3	1.315
11	17.6	1327	20.1	1.296
12	22.4	1700	19.4	1.295
13	14.8	1067	3.46	1.344
14	11.5	692.3	4.15	1.327
15	16.8	1740	2.17	1.346
16	16.8	1386	1.61	1.336
17	12.9	1149	2.20	1.339
18	16.2	1291	1.90	1.331
19	5.08	250.6	13.0	1.391
20	5.09	198.9	15.5	1.377
21	15.1	926.2	6.22	1.367
22	10.4	721.7	3.02	1.347
23	24.5	1557	5.30	1.324
24	19.2	1496	1.99	1.312
25	13.85	937.5	3.32	1.353
26	11.84	854.8	2.06	1.342
27	13.79	784.8	7.03	1.363
28	14.45	968	3.17	1.367

3.5.2 Data analysis

The data was first mean centered and unit variance scaled as described in Section 2.6. Data analysis was performed using the JMP® 9.0 statistical software package by SAS [65]. PCA was used as the projection method to analyse multivariate data and examine the overall data structure. Figure 3.5 is an outlier analysis based on Hotelling's T^2. The data contains no outliers.

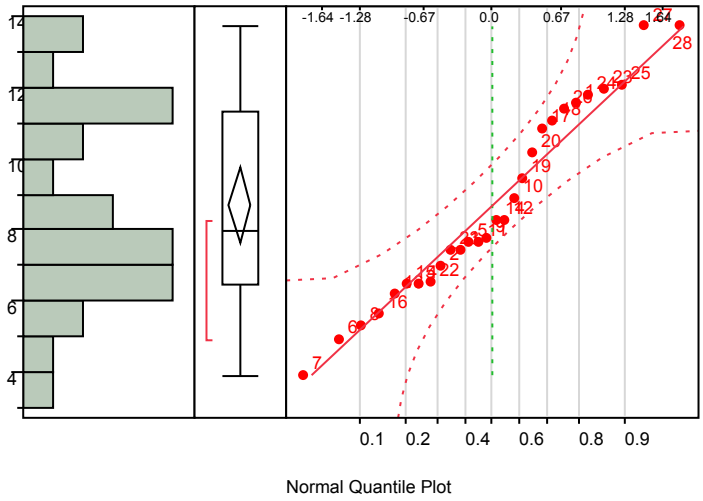


Figure 3.5: Distribution plot, outlier box plot and normal quantile plot.

Figure 3.6 is a scree plot of eigenvalues vs. number of principal components (PCs) combined with a *pareto plot*. The pareto plot shows each eigenvalue as a percentage of the total eigenvalue. Eigenvalues sum to the number of variables when the principal component analysis is done on the correlation matrix. Cumulative percent shows the cumulative percent of variation represented by the eigenvalues. The scree plot is useful for visualizing the dimensionality of the data space. In this example, the scree plot suggests that using three PCs is adequate as any additional PC did not substantially increase the amount of variance accounted for. The first three principal components (i.e. $A = 3$) that are guaranteed to be orthogonal, captured 86.82% of the total variance of $\mathbf{X}_{mix\ all}[= (\mathbf{R}_{starch} \cdot \mathbf{X}_{starch})\mathbf{R}_{oter}\mathbf{Z}]$

data. The score variables generated by PCA are optimal summaries of the original variables. Table 3.4 and Table 3.5 contain the score and the loading values, respectively, for **X**- and **Y**- blocks.

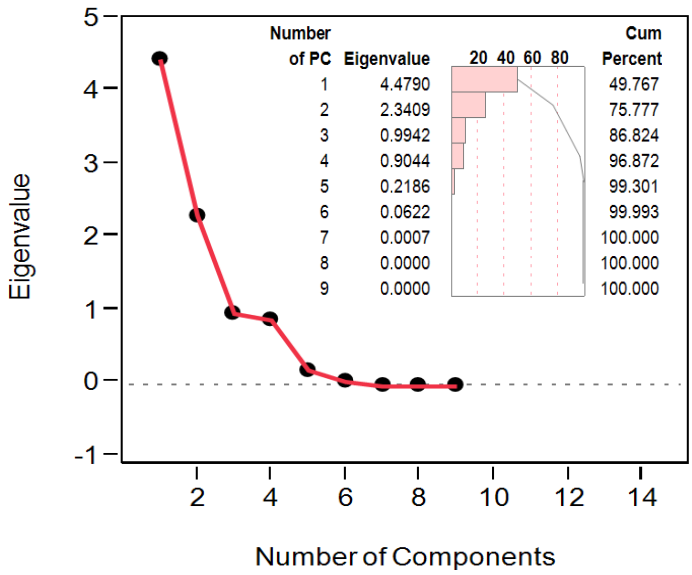


Figure 3.6: Scree plot and pareto plot for PCA on X-variables.

Figure 3.7 and Figure 3.8 present loading plots combined with score plots for the PCA on **X**-variables. There are 28 mixture observations involved in this problem. Since a three dimensional plot is the highest possible dimensional plot that can be constructed, only three variables can be studied at a time. Therefore, the relationship between all 28 observations is hidden in a highly dimensional space. The score plot projects high dimensional variables onto low dimensional variables thereby facilitating the analysis of all 28 observations simultaneously on a two dimensional plot (see Section 2.6.1.2). The loading plot shows the relationships among the 7 variables. The variables that are most influential for the model are found on the periphery whereas the less influential variables are encountered around the origin of the loading plot (see Section 2.6.1.3). Since the score and the loading are complementary and superimposable, a *biplot* can be created that combines both plots (Figure 3.7 and Figure 3.8).

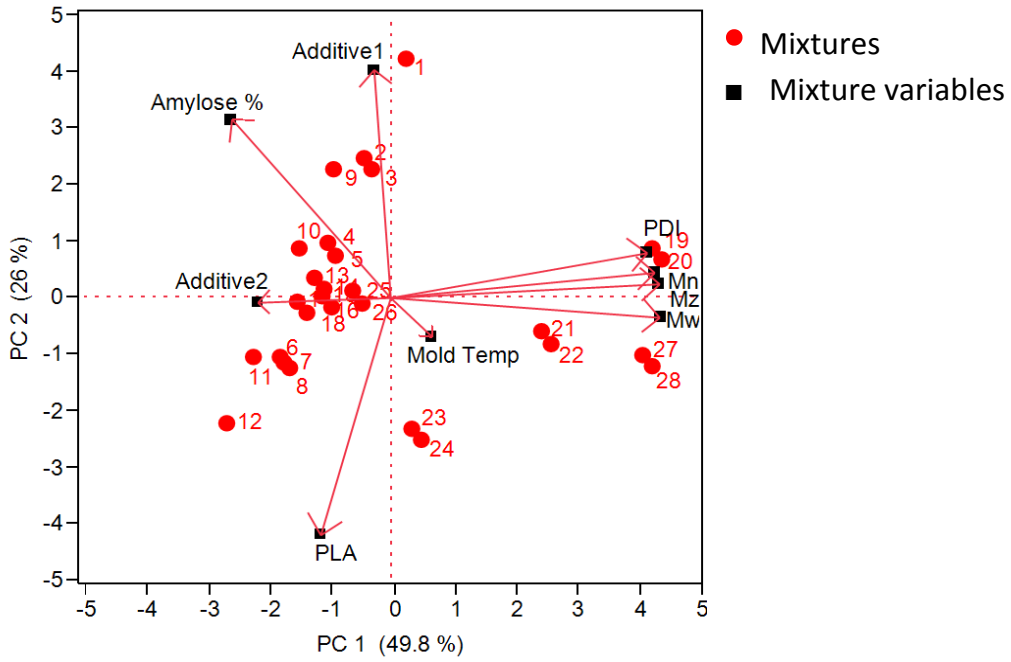


Figure 3.7: Combined PCA score and loading plots (Biplot) on first and second components for X-block.

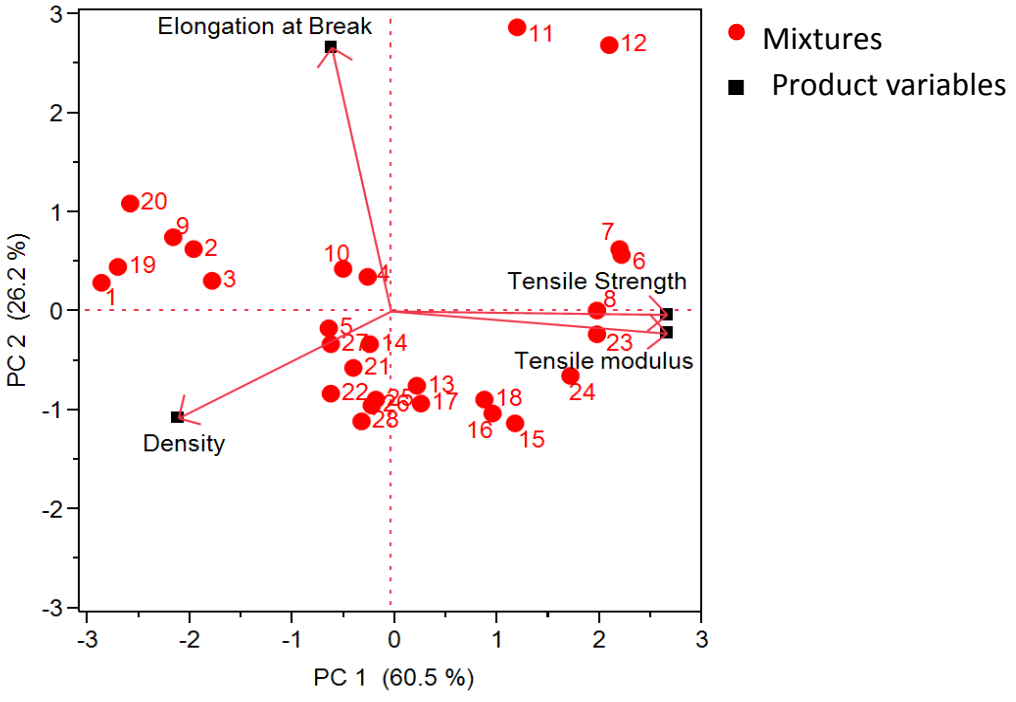


Figure 3.8: Combined PCA score and loading plots (Biplot) on first and second components for Y-block.

It is clear from Figure 3.7 that the variables *PDI*, *Mn*, *Mz* and *Mw* are positively correlated, i.e. when one increase or decreases, the other has a tendency to change in the same direction. On the other hand, the variable molding temperature is negatively or inversely correlated to amylose% i.e. increase in one will have inverse effect on the other, and vice versa. Similarly, Figure 3.8 shows a loading plot combined with a score plot for the PCA on **Y**-variables (which explained 86.79% of the variance of the final product properties). In this loading plot all product property variables, except the tensile strength and the tensile modulus, have the trade-off relationship among them.

Figure 3.9 shows the regression coefficients for the final product properties and how the starch properties, PLA and additives are correlated. For instance: high additive1, low molding temperature and low amylose content leads to high elongation. High lactic acid and low additive1 leads to high tensile- strength and modulus. The observations are well known to most experienced polymer chemists [64]. The results obtained here are consistent with Muteki's results [63]. Since only three properties can be represented in the ternary cluster diagram and tensile strength and modulus are affected similarly by the mixture factors, only one of them will be used together with elongation and density as three targeted product properties. However, in situations where more than three properties have to be considered, an algebraic approach may be used [24, 21].

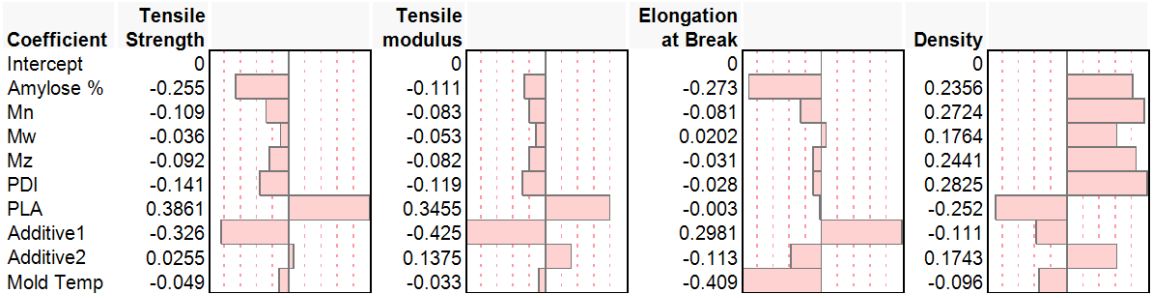


Figure 3.9: PLS model coefficients for blend property matrix Y using four latent factors.

Table 3.4: PCA score values for X- and Y-blocks.

X score value			Y score value	
t_1	t_2	t_3	t_1	t_2
0.252	4.221	-0.155	-2.814	0.288
-0.426	2.466	-0.204	-1.928	0.625
-0.279	2.257	1.655	-1.743	0.299
-1.022	0.948	-0.248	-0.215	0.330
-0.874	0.740	1.611	-0.597	-0.179
-1.784	-1.043	-0.302	2.264	0.566
-1.711	-1.148	0.628	2.248	0.618
-1.637	-1.252	1.557	2.023	0.002
-0.922	2.265	-0.817	-2.117	0.737
-1.459	0.858	-0.854	-0.463	0.413
-2.204	-1.049	-0.909	1.237	2.864
-2.657	-2.219	-0.942	2.147	2.688
-1.217	0.342	-0.952	0.259	-0.770
-1.070	0.133	0.908	-0.196	-0.346
-1.093	0.018	-1.011	1.222	-1.133
-0.946	-0.191	0.848	1.002	-1.043
-1.507	-0.072	-1.494	0.301	-0.936
-1.359	-0.280	0.365	0.918	-0.905
4.256	0.867	-1.106	-2.659	0.435
4.403	0.658	0.753	-2.535	1.090
2.462	-0.608	-0.949	-0.355	-0.572
2.609	-0.816	0.910	-0.588	-0.832
0.369	-2.330	-0.767	2.025	-0.247
0.516	-2.539	1.093	1.765	-0.659
-0.602	0.111	-0.408	-0.134	-0.900
-0.455	-0.098	1.451	-0.188	-0.958
4.105	-1.016	-1.259	-0.590	-0.347
4.252	-1.224	0.601	-0.287	-1.128

Table 3.5: PCA loading values for X- and Y-blocks.

X loading value			Y loading value	
p_1	p_2	p_3	p_1	p_2
-0.276	0.469	0.097	0.616	-0.012
0.460	0.066	-0.069	-0.132	0.924
0.468	-0.052	-0.078	-0.474	-0.374
0.468	0.036	-0.067		
0.447	0.118	-0.043		
-0.122	-0.623	0.008		
-0.029	0.599	0.095		
-0.231	-0.013	-0.359		
0.072	-0.102	0.914		

3.5.3 Model development

Latent variable models were developed using three latent variables (t_i, obtained from PCA) to predict three product properties: tensile strength (*TS*), elongation at break (*EB*) and density (*Rho*). The general form of the equation is:

$$\hat{y} = \beta_0 + \sum_{i=1}^{3} \beta_i t_i + \sum_{i<j}^{3} \sum_{j>i}^{3} \beta_{ij} t_i t_j + \sum_{i=1}^{3} \beta_{ii} t_i^2 \qquad \textbf{(3.11)}$$

Since addition of parameters result in the risk of overfitting, it is important to select the optimum number of parameters during model selection. *Bayesian Information Criterion* (*BIC*) introduces a penalty term for the number of parameters chosen in the model. *BIC* can be expressed as:

$$BIC = -2\log likelihood + B\ln(M) \qquad \textbf{(3.12)}$$

where, B = the number of parameters, including intercept and error terms in the model

M = the number of observations in the data set.

Minimization of *BIC* was used as a criterion during the model selection process in the statistical package JMP 9.0 and the detailed expressions for the property models are as follows:

$$TS = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \beta_{13} t_1 \cdot t_3 + \beta_{33} t_3^2 \quad \textbf{(3.13)}$$

$$EB = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \beta_{12} t_1 \cdot t_2 + \beta_{23} t_2 \cdot t_3 + \beta_{11} t_1^2 + \beta_{22} t_2^2 \quad \textbf{(3.14)}$$

$$Rho = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_3 t_3 + \beta_{12} t_1 \cdot t_2 + \beta_{11} t_1^2 + \beta_{22} t_2^2 \quad \textbf{(3.15)}$$

The estimated regression coefficients for the respective property models are listed in Table 3.6. Figure 3.10 shows a plot with actual versus predicted properties.

Table 3.6: The regression coefficients for the expressions in Eqs. (3.13) through Eqs. (3.15).

Regression coefficients	*TS* (MPa)	*EB* (%)	*Rho* (g/cm³)
β_0	14.59	2.534	1.345
β_1	-1.119	-1.328	0.0146
β_2	-3.938	0.4816	0.0073
β_3	-0.1558	-0.7965	-0.0068
β_{12}	-	0.7488	0.0016
β_{23}	-	0.6859	-
β_{13}	-0.4749	-	-
β_{11}	-	0.7361	-0.0018
β_{22}	-	0.5084	-0.0013
β_{33}	-1.59	-	-
R^2	0.917	0.698	0.852
R^2_{adj}	0.898	0.593	0.810

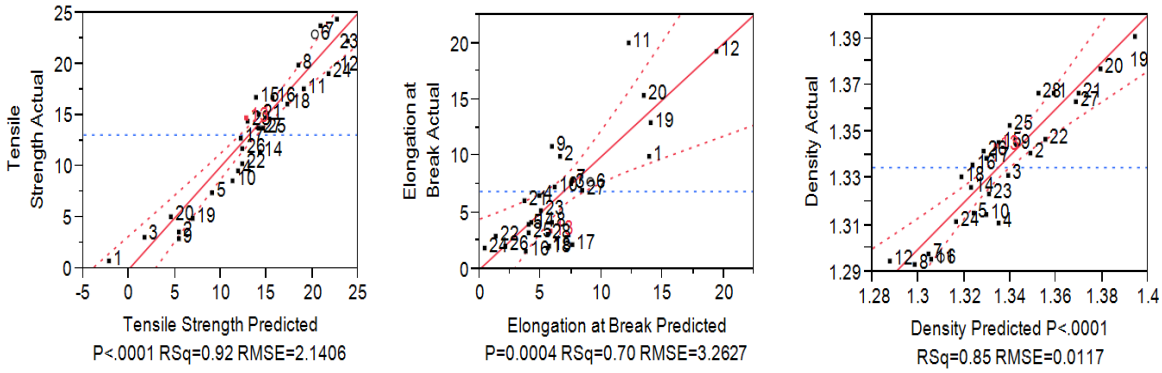


Figure 3.10: Predicted vs. actual product properties using PCR model.

3.5.4 Design of desired products in score space

It is desirable to cover a wide range of product properties with a minimum number of products in order to minimize the manufacturing costs, inventories and material cost. Figure 3.11 is a product properties score plot (same as Figure 3.8) where the desired product properties can be independently selected such that a set of product grades spans the desired property space [63]. Here five target products are selected in the score space to provide a wide range of final product properties. Table 3.7 shows the score values from the selected points in Figure 3.11. Using the score and the loading values from the PCA on **Y** (Table 3.5), the properties for the desired products (tabulated in Table 3.8) can be calculated as

$$\hat{y}_{des} = 1 \cdot \bar{x}_h + \left(\sum_{a=1}^{2} t_{des,a} \cdot p_a^T \right) \cdot s_h \qquad \textbf{(3.16)}$$

where, $1 \cdot \bar{x}$ = the variable average

s = the standard deviation originated from the pre-processing step.

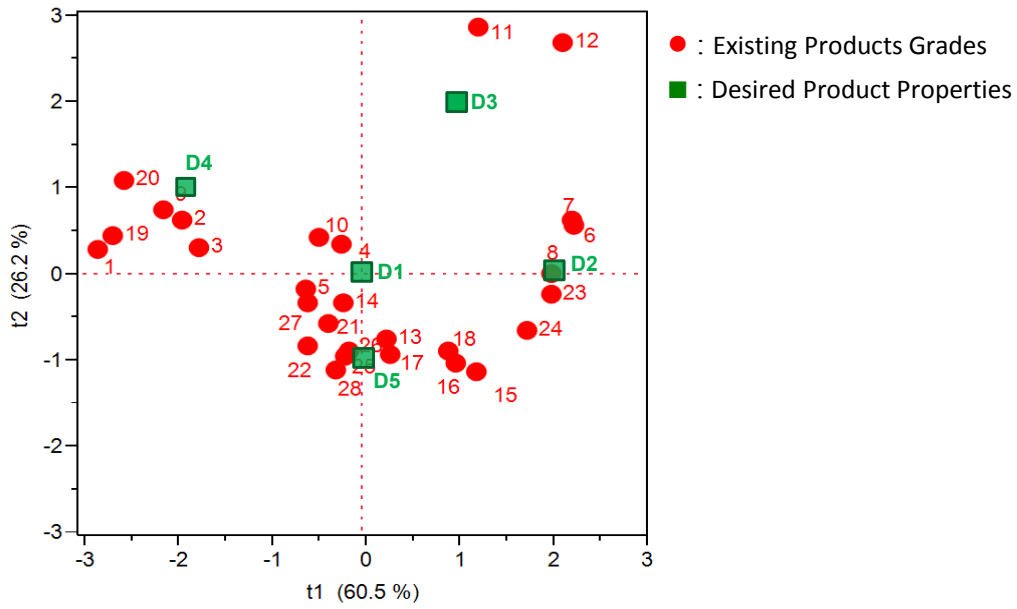


Figure 3.11: Visualization of target product properties in score space.

Table 3.7: Score values for desired products in Figure 3.11.

Products	t_1	t_2
D1	0.0	0.0
D2	2.0	0.0
D3	1.0	2.0
D4	-2.0	1.0
D5	0.0	-1.0

Table 3.8: Desired product properties for desired products in Figure 3.11.

Products	*TS* (MPa)	*EB* (%)	*Rho* (g/cm³)
D1	13.06	6.861	1.335
D2	21.32	5.510	1.309
D3	17.03	15.64	1.302
D4	4.725	12.94	1.350
D5	13.14	2.136	1.345

From the five desired products, product D2 (from Table 3.8) was selected as a target product (♦) in score space. The mixture conditions (i.e., the raw materials, their blend ratios, and process condition) that give the desired product properties of the target product can be estimated by minimizing the sum of square differences between the target and predicted properties. The results are listed in Table 3.9. Using this information, a new blend product can be synthesized.

Table 3.9: Required mixture conditions to achieve target product properties.

Target	R_{starch}			R_{other}			Z
	Starch 1	Starch 2	Starch 3	PLA	Add 1	Add 2	Temp ^{o}C
D2	0.171	0.132	0.053	0.404	0.223	0.016	159.5

Muteki used a mixed integer non-linear programming (MINLP) algorithm to find raw materials that were not part of previous product grades. The results can be found in [63]. Such optimization is outside of the scope of this study.

3.5.5 Design of desired products in cluster space

The ternary cluster space provides an excellent platform for simultaneous visualization and solution of mixture design problems. Figure 3.12 is a ternary cluster plot with a feasibility region for blend product properties that incorporates all the existing product grades and all five desired product properties identified from Figure 3.11 and Table 3.8. The method to convert score values to cluster values is discussed in Section 2.7. The mixture feasibility region in cluster space ensures that the products within its boundary are physically feasible, consistent with past operating strategies and expected to yield the desired product qualities.

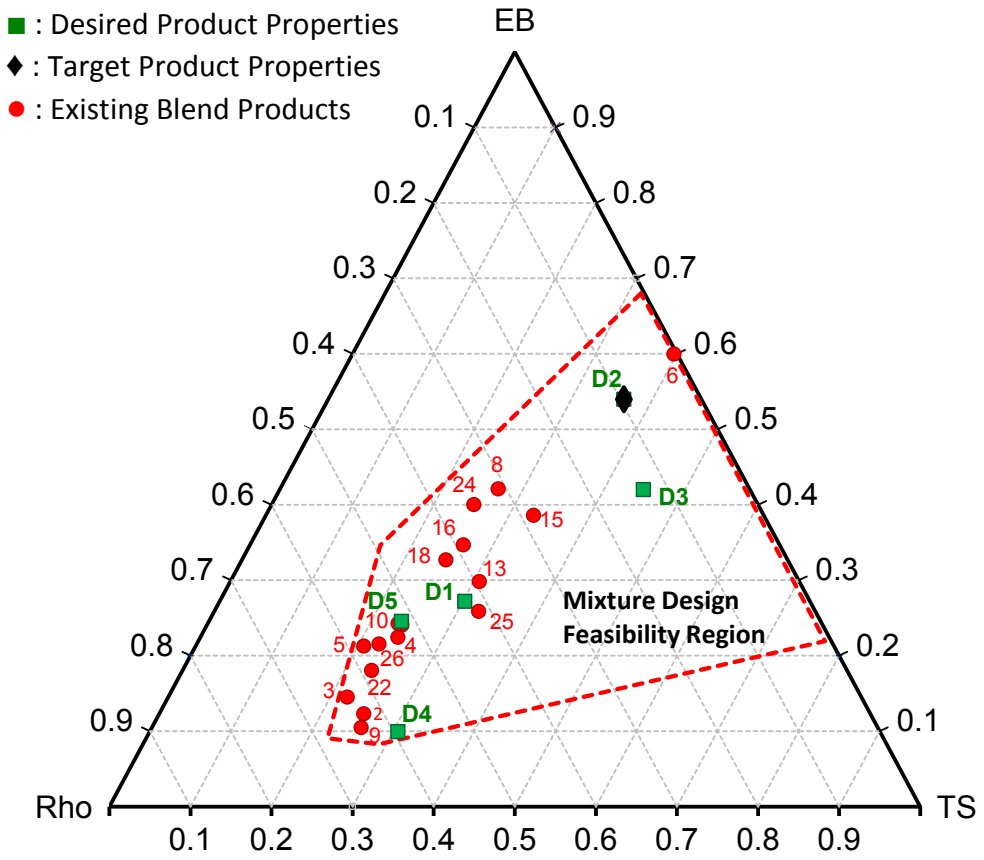


Figure 3.12: Visualization of starch blending formulation in cluster space.

Analogous to the use of the score plot to identify a product whose properties span the desired property space, cluster space can achieve similar objectives. Cartesian coordinate points for D2 cluster are (0.634, 0.540). Since property clusters are tailored to maintain the fundamental rules for intra- and inter- stream conservation, the mixing operation can be optimized using lever-arm analysis. For instance, if two polymer product grades are compatible, i.e. miscible, they can be mixed together to get a polymer with properties somewhere between those of the two polymers mixed. In latent property cluster space, one can rapidly screen out inherently infeasible combinations of candidate constituents visually. The reduced search space can then be explored for the feasibility of formulating binary, ternary and

multi-component mixtures to achieve optimum products based on lever-arm analysis without extensive enumeration.

For example, the target product (D2) synthesized in Table 3.9 can also be formulated by mixing existing blends 24 and 25 with blend 6 (denoted as 24-25-6 in Table 3.10). In Figure 12, M1 represents a mixture of existing blend products 24 and 25 in a proportion 0.640 and 0.360 respectively. When M1 is mixed with blend 6 in a proportion 0.207 and 0.793 respectively, product target (D2) with property targets (listed in Table 3.8) can be achieved. A simple visual analysis of Figure 3.13 suggests that the binary mixture of existing blend products is inherently infeasible to achieve the target product.

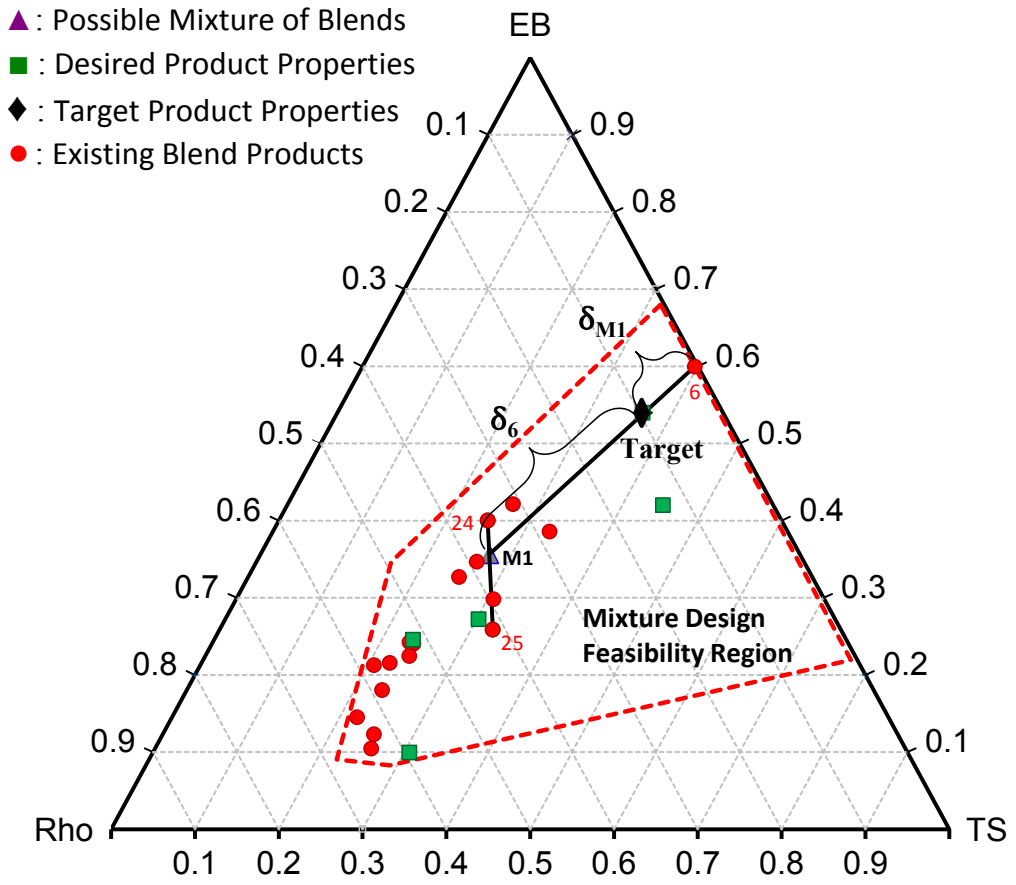


Figure 3.13: Visualization of starch blending formulation in cluster spac

Table 3.10: Candidate ternary mixtures and fractional contributions of the constituents.

S.N.	Mixture (A-B-C)	Feasible	A	B	C
1	8-2-6	Yes	15%	5%	80%
2	8-3-6	No	12%	6%	82%
3	8-4-6	No	10%	10%	80%
4	8-5-6	No	6%	11%	83%
5	8-9-6	Yes	16%	5%	79%
6	8-10-6	No	11%	8%	81%
7	8-13-6	Yes	15%	8%	77%
8	8-15-6	Yes	14%	13%	73%
9	8-16-6	No	1%	21%	78%
10	8-18-6	No	1%	20%	79%
11	8-22-6	No	12%	7%	81%
12	8-25-6	Yes	16%	7%	77%
13	8-26-6	No	9%	9%	82%
14	24-2-6	Yes	12%	6%	82%
15	24-3-6	No	10%	7%	83%
16	24-4-6	No	8%	10%	82%
17	24-5-6	No	5%	12%	83%
18	24-9-6	Yes	14%	6%	80%
19	24-10-6	No	9%	9%	82%
20	24-13-6	Yes	12%	10%	78%
21	24-15-6	No	12%	14%	74%
22	24-16-6	No	1%	23%	76%
23	24-18-6	No	1%	21%	78%
24	24-22-6	No	10%	8%	82%
25	24-25-6	Yes	13%	8%	79%
26	24-26-6	No	7%	10%	83%

However, a total of 26 candidate ternary mixtures were identified. In order to validate the feasibility of the designed formulation, the *AUP* values of the formulated mixtures and the target must match along with matching the cluster targets. After performing this analysis, only 9 out of 26 identified

candidate ternary mixtures matched the cluster target and *AUP* value of the target product and the results are summarized in Table 3.10. The 17 infeasible mixtures, although, matched the cluster targets but had *AUP* values higher than ±1% of the target value (1.802). Therefore, the discrepancy was assumed to be higher than the accuracy that can be expected by graphical lever-arm analysis.

3.6 Conclusion

This research effort has focused on the formulation and solution of product design problems by systematic and insightful use of past data. Properties of the raw materials, their blend ratios, and the process conditions are used to predict and enhance the performance of a target product. Using the duality of linear programming to solve the design problem in the lower dimensional property domain, instead of high dimensional component space, significantly reduced the computational complexity of the problem. Moreover, the ternary diagram provides a quick targeting tool that aides in the evaluation, analysis, and screening of alternatives. The approach differs from conventional techniques because it is non-iterative, avoids the combinatorial explosion when multiple components are involved, and avoids the difficulty of formulating and solving the mixed integer non-linear programming involved in many mixture design problems. Such a practice in industry results in fewer trials and experiments to run, thereby saving resources, capital and most importantly the product development time.

However, the method and example presented above is limited to selecting, i.e. identifying candidates from a database of known raw materials. The raw materials used, are selected from a list of pre-defined candidate components, therefore limiting the performance to those components. The problem here is that these decisions are made ahead of design and are usually based on qualitative (or at best quantitative) process knowledge

and/or experience and thus possibly yield a sub-optimal design. In order to guarantee global optimality, all possible compounds must be considered. Chapter 4 will present a biodiesel additive design problem to demonstrate molecular design concepts.

CHAPTER 4

DESIGN OF BIODIESEL ADDITIVES

4.1 Introduction

High prices and environmental impact related to fossil-based raw materials are the driving force for sustainable development concepts and their use in industry. This will also have implications for the design of chemical products and their production routes. Biofuels, mainly ethanol used in gasoline engines and *fatty acid alkyl esters* (biodiesel) as well as their blends with petro-diesel used in diesel engines, are some of the few alternatives that have not required significant new infrastructure or change on the part of consumers or auto manufacturers. As a result, these biofuels are the most widely deployed substitute for conventional fossil fuels in transportation today. In the U.S., biofuels can currently be blended up to 10% (ethanol) and 20% (biodiesel) in every gallon of fossil fuel. Among these alternatives, biodiesel has become a fast growing market and is expected to outpace gasoline demand [66, 67, 68].

Some advantages of biodiesel, compared to petro-diesel, include:

- renewability and domestic origin,
- biodegradability and sustainable,
- higher flash point,
- reduction of most regulated exhaust emissions due to lack of sulfur,
- miscibility in all ratios with petro-diesel,
- compatibility with the existing fuel distribution infrastructure, and
- inherent lubricity.

Any type of feedstock that contains free fatty acids and/or triglycerides such as vegetable oil, waste oil, animal fat, and waste grease can be converted into biodiesel. The American Society for Testing and Materials (ASTM) defines biodiesel as a fuel comprised exclusively of monoalkyl esters of long-chain fatty acids derived from vegetable oils or animal fats, designated B100 (100% pure biodiesel), meeting the requirements of ASTM designation D6751 [69].

Biodiesel is a very good example of chemical product design. A high viscosity and high melting point make its use directly as a fuel in common engines difficult. The conversion of triglycerides into methyl or ethyl esters through the *transesterification* (also called alcoholysis) process reduces the molecular weight to one-third that of the triglyceride, reduces the viscosity by a factor of about eight and marginally increases the volatility [70]. Transesterification is an ester conversion process that splits up the triglyceride (TG); that is, it takes the glycerol (GL) of the TG and replaces it with alkyl radical of the alcohol used [71]. In Figure 4.1, R_1, R_2, and R_3 represent long chain fatty acid radicals of the mixed TG used whereas R' represents an alkyl radical of the alcohol used. Some common fatty acids, which exist in the triglyceride molecule are shown in Table 4.1.

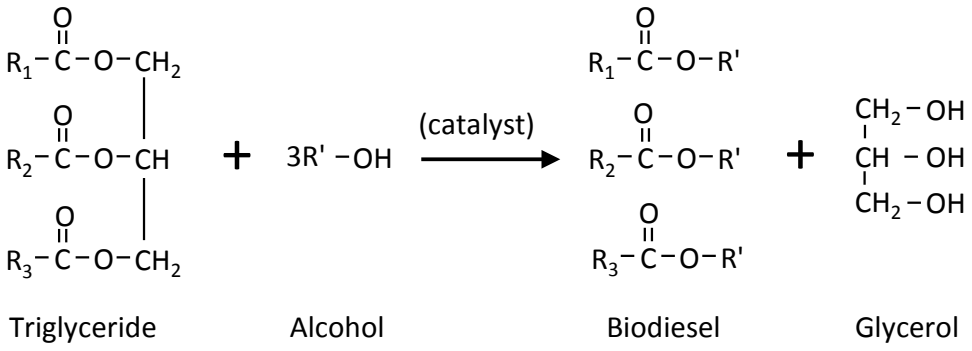


Figure 4.1: Overall stoichiometric transesterification reaction scheme.

Glycerol is the main byproduct in biodiesel production; almost 13% of glycerol comes from biodiesel production. Therefore, a valuable use of glycerol

is a very important success factor for this technology. Product design principles have to be applied to search for profitable applications that also fulfill the requirements of the sustainability principles [72].

Table 4.1: Chemical structures of common fatty acids.

Fatty Acid	CXX:Y	Chemical Structure
Myristic acid	(C14:0)	CH_3-$(CH_2)_{12}$-COOH
Palmitic acid	(C16:0)	CH_3-$(CH_2)_{14}$-COOH
Stearic acid	(C18:0)	CH_3-$(CH_2)_{16}$-COOH
Oleic acid	(C18:1)	CH_3-$(CH_2)_7$-CH=CH-$(CH_2)_7$-COOH
Linoleic acid	(C18:2)	CH_3-$(CH_2)_4$-CH=CH-CH_2-CH=CH-$(CH_2)_7$-COOH
Linolenic acid	(C18:3)	CH_3-CH_2-CH=CH-CH_2-CH=CH-CH_2-CH=CH-$(CH_2)_7$-COOH
Arachidic acid	(C20:0)	CH_3-$(CH_2)_{18}$-COOH
Behenic acid	(C22:0)	CH_3-$(CH_2)_{20}$-COOH
Erucic acid	(C22:1)	CH_3-$(CH_2)_7$-CH=CH-$(CH_2)_{11}$-COOH

If methanol is used, the biodiesel produced is *fatty acid methyl ester* (FAME). For example, the structure of stearic acid methyl ester can be obtained by replacing the H atom in the COOH- group with a CH_3 group as shown in Figure 4.2.

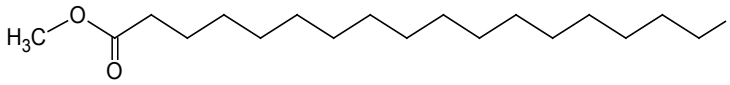


Figure 4.2: Stearic acid methyl ester.

Figure 4.3 shows a typical triglyceride structure of soybean oil that is made up of mixed fatty acid fragments and a glycerol fragment.

The physical and chemical fuel properties of biodiesel basically depend on the fatty acids distribution of the triglyceride used in the production. Fatty acids vary in their carbon chain length and in the number of double bonds (unsaturation level), and are represented by C XX:Y where 'XX' is the number

of carbon atoms and 'Y' is the number of double bonds. The fatty acid distributions of some feedstock commonly used in biodiesel production are shown in Table 4.2 [73].

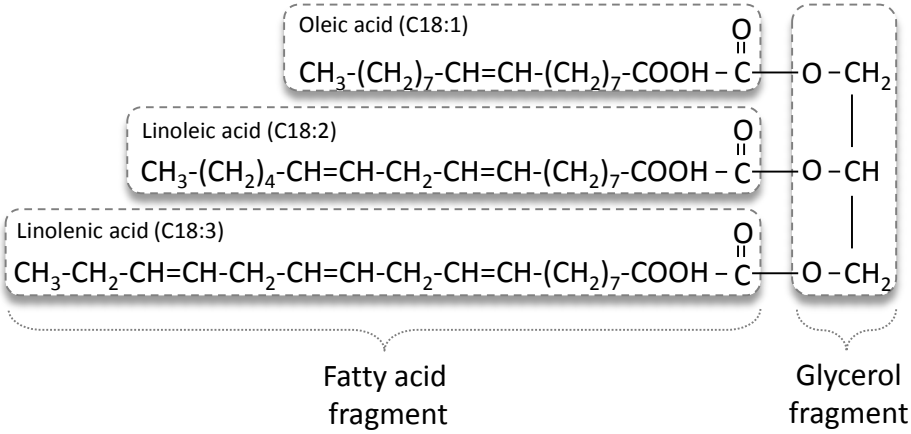


Figure 4.3: A typical triglyceride molecule with different fatty acid chains of soybean oil.

Since the fatty acid profile varies with different feedstock, the final quality of the biodiesel varies depending upon the parent feedstock used. A feedstock dependent fuel property is one of the technical difficulties associated with biodiesel that have limited its wide usability.

Table 4.2: Fatty acid profiles of some common biodiesel feedstock

	Fatty acids (wt %)						
Feedstock	**C14:0**	**C16:0**	**C16:1**	**C18:0**	**C18:1**	**C18:2**	**C18:3**
Sunflower	–	6.08	–	3.26	16.93	73.73	–
Rapeseed	–	3.49	–	0.85	64.40	22.30	8.23
Soybean	–	10.58	–	4.76	22.52	52.34	8.19
Soybean soapstock	–	17.20	–	4.4	15.7	55.6	7.10
Used frying oil	–	12.00	–	–	53.00	33.00	1.00
Tallow	3–6	24–32	–	20–25	37–43	2–3	–
Lard	1–2	28–30	–	12–18	4–50	7–13	–
Yellow grease	2.43	23.24	3.79	12.96	44.32	6.97	0.67
Brown grease	1.66	22.83	3.13	12.54	42.36	12.09	0.82

Transesterification does not alter the fatty acid composition of the feedstocks and this composition plays an important role in some critical parameters of biodiesel. The vegetable oils are mainly characterized by certain fuel related properties. Some of them are tabulated in Table 4.3 [74].

Table 4.3: Fuel properties of biodiesel fuels and diesel

Vegetable oil	Kinematic viscosity at 38°C (mm^2/s)	Cetane No. (°C)	Heating Value (MJ/kg)	Cloud Point (°C)	Pour Point (°C)	Flash Point (°C)	Density (kg/l)	Carbon residue (wt.%)
Sunflower	34.4	36.7	39.6	7.2	-15.0	274	0.916	0.27
Rapeseed	37.3	37.5	39.7	-3.9	-31.7	246	0.912	0.30
Soybean	33.1	38.1	39.6	-3.9	-12.2	254	0.914	0.25
Peanut	40.0	34.6	39.8	12.8	-6.7	271	0.903	0.24
Palm	39.6	42.0	–	31.0	–	267	0.918	0.23
Cottonseed	33.7	33.7	39.5	1.7	-15.0	234	0.915	0.24
Corn	35.1	37.5	39.5	-1.1	-40.0	277	0.909	0.24
Diesel	2.0-4.5	51.0	43.8	– 18	-25	55	0.820-0.860	–

4.2 Structure Property Relationships

Chain length and number, position and configuration of double bonds account for the variation in physical properties of fatty acids. Saturated chains are highly flexible but the fully extended conformation is the most stable because of the lack of steric interference. The tetrahedral bond angle on carbon results in a molecular geometry for saturated fatty acids that is relatively linear (see Figure 4.2). This molecular structure allows a close arrangement of the fatty acid molecules with strong intermolecular interactions.

On the other hand, the introduction of double bonds with predominantly *cis*-configuration in the hydrocarbon chain in unsaturated fatty acids results in bends of about 30 degrees in the molecular geometry (see Figure 4.4). The molecules do not arrange very closely due to chain branching. It allows more flexibility and weaker van der Waals force between the molecules.

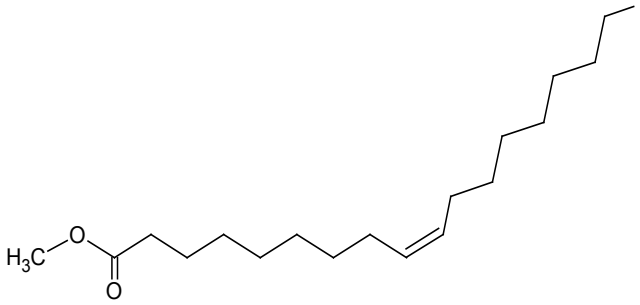


Figure 4.4: Oleic acid methyl ester.

4.3 Technical Difficulties with Biodiesel Use

Although, biodiesel, because of its biological origin, has many advantages compared to its petroleum counterparts, it also has several technical problems that have persisted to the present and have impaired its use and commercialization. Three major limitations are [71]:

- **Oxidative stability***:* Biodiesel undergoes oxidative degradation over time, mainly influenced by temperature and oxygen availability/exposure. Residual products of biodiesel such as insoluble gums, organic acids, and aldehydes formed from the degradation may cause engine deposits and injection problems. Biodiesel is essentially non-aromatic. Petroleum diesel contains essentially no olefinic bonds, while biodiesel can contain a significant number of these reactive, unsaturated sites that provide pathways for oxidation instability.
- **Low-temperature Operability:** Biodiesel contains wax molecules, which are dissolved in the fuel at higher temperatures. As the fuel temperature drops, the wax molecules begin to crystallize. At lower temperature, the larger crystals fuse together and form agglomerations that eventually prevent pouring of the fuel and plugging of filters.
- **Feedstock dependent fuel property:** The fatty acid profile of biodiesel is identical to that of the parent oil or fat. Therefore, feedstock origin will impact the final quality of the biodiesel product. To produce fuel grade biodiesel, the characteristics of the feedstock are

very important during the initial research and production stage (fuel properties of diesel and biodiesel fuels from various sources are given in Table 4.3.)

Moreover, simultaneous solution of these problems has proven difficult as improvements in one area tend to impair another. Due to the inverse relationship between oxidative stability and low-temperature operability, the design of an optimal fuel for all environments can be a rather difficult task. Structural factors that improve oxidative stability adversely influence low-temperature operability and vice versa. Figure 4.5 is a graphical representation of the effects of the biodiesel feedstock profile on the biodiesel properties published by the National Renewable Energy Laboratory in 2007.

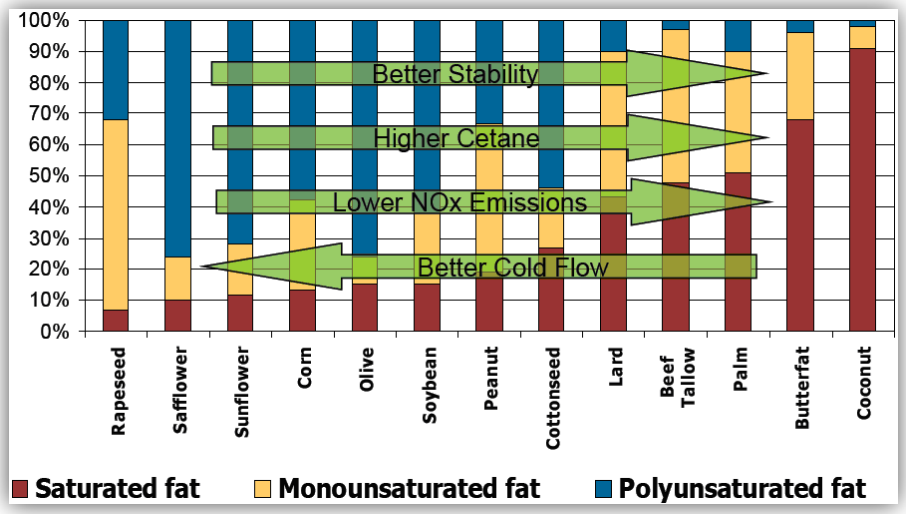


Figure 4.5: Compositions of fats and oils and their effects on the fuel properties.

To ensure a uniform quality of biodiesel produced from vegetable oils or animal fats, the final products must meet stringent international and regional quality requirements such as ASTM D 6751 [69] in the U.S. and EN 14214 [75] in Europe. These standards identify the parameters the pure biodiesel (B100) must meet before being used as a pure fuel or being blended with petroleum-based diesel fuel. Table 4.4 provides the specifications for biodiesel and diesel.

Table 4.4: Specifications for biodiesel and diesel

Specification		EN 14214:2008	ASTM D 6751:2009	EN 590:1999
Applies to	**Units**	**FAME[a]**	**FAAE[b]**	**Diesel**
Kinematic Viscosity	mm^2/s, @ 40°C	3.5-5.0	1.9-6.0	2.0-4.5
Cetane number	-	51 minimum	47 minimum	51 minimum
Cloud point	°C	-	report	-
CFPP	°C	Location and time specific	-	Location and time specific
Oxidation stability	hr, @110°C	6 minimum	3 minimum	N/A (25 g/m^3)

[a] refers to fatty acid alkyl esters
[b] refers to fatty acid methyl esters

The listed specifications in biodiesel standards are directly influenced by the fatty acid profile of the biodiesel fuel which is directly influenced by the fatty acid profile of parent oil used (see Table 4.1 and Table 4.2).

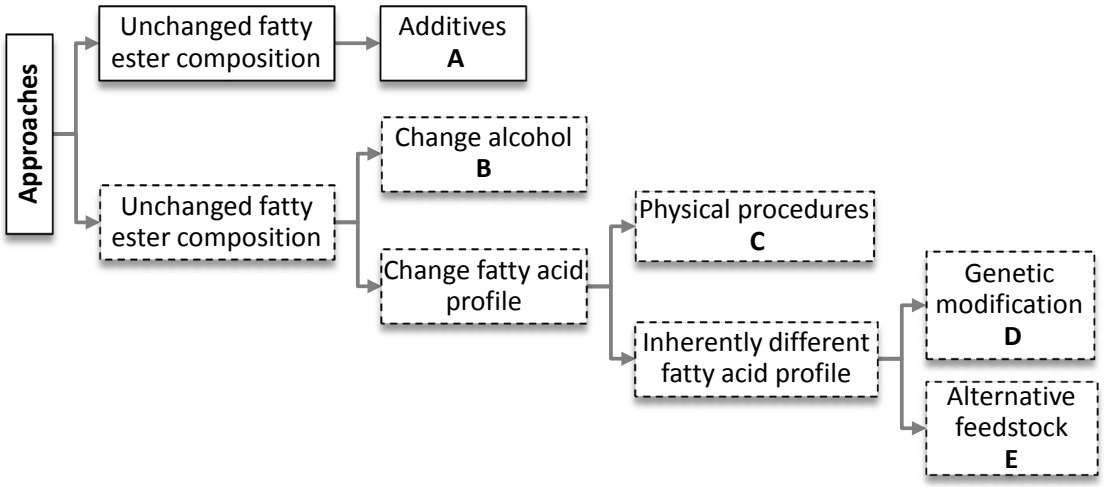


Figure 4.6: Approaches to improving biodiesel fuel properties.

Several pathways are possible for improving the fuel-related properties of biodiesel. Figure 4.6 presents an overview of the various approaches (A to E) that have been explored [76, 77].

4.4 Fuel Additives

Fuel additives, such as antioxidants, cetane enhancers, or cold-flow improvers have become a common and indispensable tool not only to decrease the drawbacks described above but also to produce specified products that meet international and regional standards, allowing fuels trading to take place. However, an additive solution to one problem often aggravates another problem. In addition, the questions of additive compatibility, required addition levels, the effect on other properties, and whether these additives function as designed for biodiesel fuels with differing fatty acid profiles still remain challenges that require further investigation [78].

Therefore, it is desired to molecularly design biodiesel additives to account for the unintended effect on other fuel properties in the neat and the blend fuel in order to achieve the performance properties of the petroleum-based fuel. In this way, biofuels can be formulated that are adaptable to a range or blend of feedstocks and the desirable fuel characteristics like oxidative stability and wide operating temperature range. Moreover, such a sustainable biofuel must also meet the specifications required by the transportation and the aviation industry.

4.5 Additive Design

Fuel characterization data obtained with near infrared spectroscopy (see Section 2.4) is combined with property clustering techniques (see Section 2.2.2) in a reverse problem formulation (see Section 2.2.1) to design additive molecules which, when mixed with off-spec biodiesel, produce biodiesel that meets the desired fuel specifications. The characterization data consists of a multitude of property values (such as cetane number, melting point, and kinematic viscosity) specified by the aviation industry to ensure adequate performance. To facilitate an efficient design we propose consolidating these various properties into a latent property domain using principal component

analysis (PCA) and principal component regression (PCR) techniques (see Section 2.6.1 and 2.6.2). Characterization-based molecular design using group contribution parameters are then used to build novel additives that match the fuel specifications in the latent property space. Sustainability may be controlled by environmental and ecological constraints on the design of the additives. The additives found can then be used to offset the impact of the feedstock residuals on the biofuel blend properties.

4.5.1 Types of Additives

Different additives that are commercially available to improve diesel fuel performance are selected as the training set molecules [79, 80]. Antioxidant additives can help slow the degradation process and improve fuel stability. Cold flow enhancers can improve the cold-flow properties to solve the low temperature operability problem. Cetane improvers can help improve ignition properties thereby reducing NOx emission. Table 4.5 shows some of the available additives used as training set in this formulation study.

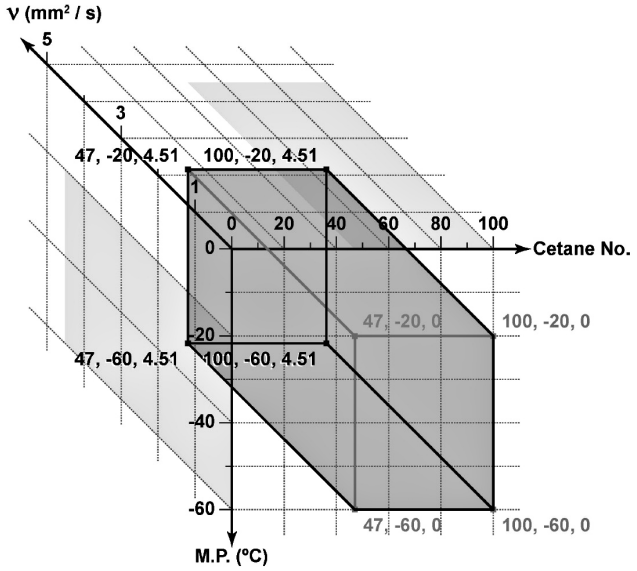


Figure 4.7. Target specifications for bio-diesel and its blend in terms of cetane number, melting point, and kinematic viscosity.

Since cetane number, cold flow, kinematic viscosity, and oxidation stability are critical properties for the operation of a fuel in a diesel engine, Knothe [78] proposed a three-dimensional plot of the cetane number, melting point, and kinematic viscosity to describe the optimum properties to meet the ASTM D6751 fuel requirements. Figure 4.7 visually represents this required biodiesel property space by the shaded rectangular box [78].

4.5.2 Additive Property Estimation

Since many additives are proprietary products, their physical-chemical properties are not readily available and therefore experimental data such as melting point temperature, viscosity and cetane number are difficult to find. Therefore, for the commercially available additive compounds, these properties are estimated using structural information alone. The estimated values are later used as the measured values for property model calibration. However, the property values used in the calibration would normally be measured in an industrial setup.

***Melting temperature*:** The normal melting point temperature (T_m, K) values were estimated using a group contribution expression [34]:

$$\exp\left(\frac{T_m}{t_{mo}}\right) = \sum_i N_i T_{m1i} + \sum_j M_j T_{m2j} + \sum_k O_k T_{m3k} \qquad \textbf{(4.1)}$$

where, N_i, M_j and O_k are the number of first-, second-, and third-order groups of types i, j and k, respectively.

***Viscosity*:** The dynamic viscosity (η, mPa.s) values at 300 K were estimated using the GC+ method. GC+ combines the group contribution (GC) method and the atom-connectivity indices (CI) method [81].

$$\ln(\eta) = \underbrace{\sum_i N_i \eta_{1i} + \sum_j M_j \eta_{2j} + \sum_k O_k \eta_{3k}}_{\text{GC Terms}} + \underbrace{F\left(\eta^*\right)}_{\text{CI Terms}} \qquad \textbf{(4.2)}$$

where, $F(\eta^*)$ is a function of viscosity for all missing GC groups/fragments.

$$F(\eta^*) = \sum_{k=1}^{K} n_k F(\eta)_k + d \qquad \textbf{(4.3)}$$

where, $F(\eta)_k$ = a function of the viscosity contribution for a missing GC group/fragment k,

K = the number of missing groups/fragments,

n = the number of times a missing group/fragment appears in the molecule, and

d = a constant [81].

$$F(\eta) = \sum_i a_i A_i + b\left({}^{\upsilon}\chi^0\right) + 2c\left({}^{\upsilon}\chi^1\right) \qquad \textbf{(4.4)}$$

where, A_i = the occurrences of the i^{th} atom in the molecular structure,

a_i = the contribution of atom i, and

b & c = adjustable parameters.

The zero-order (atomic) connectivity index (${}^{\upsilon}\chi^0$) and the first-order (bond) connectivity index (${}^{\upsilon}\chi^1$) are defined by [82]

$$ {}^{\upsilon}\chi^0 = \sum_{i=1}^{L}\left(\frac{1}{\sqrt{\delta_i^{\upsilon}}}\right) \qquad {}^{\upsilon}\chi^1 = \sum_{i=1}^{M}\left(\frac{1}{\sqrt{\beta_i^k}}\right) \qquad \text{where, } \beta^k = \delta_i^{\upsilon}\delta_j^{\upsilon} \qquad \textbf{(4.5)}$$

where, L = the number of atoms in the hydrogen suppressed graph,

M = the number of bonds in the graph, and

β^k = the bond indices defined by atomic indices δ^{υ} (the values can be found in [81]).

The kinematic viscosity can be converted into dynamic viscosity through the density.

***Cetane number*:** The correlation used for estimation of the cetane number of additives is from Lapuerta *et al.* [83]

$$CN = -21.157 + \left(7.965 - 1.785N_{db} + 0.235N_{db}^2\right)N_c - 0.099N_c^2 \quad \textbf{(4.6)}$$

where, N_{db} = the number of double bonds and

N_c = the number of carbon atoms in the molecule.

Table 4.5 tabulates the estimated properties for different type of additives.

Table 4.5: Commercially available diesel additives and their estimated properties.

Type	Compound		$exp(T_m/T_{mo})$	$ln(\eta)$	CN
Oxidative stability improvers	Butylated hydroxytoluene	(BHT)	10.11	0.465	76.04
	t-butylhydroquinone	(TBHQ)	13.26	0.465	48.59
	Isopropyl alcohol	(IPA)	4.114	0.465	1.847
	Pyrogallol	(PY)	18.791	0.465	14.79
	Methyl tert-butyl ether	(MTBE)	3.484	0.465	16.19
Cold flow improvers	Polymethyl Methacrylate	(PMMA)	3.035	-1.062	15.19
	Ethylene glycol methyl ether acrylate	(EGMEA)	4.299	-0.517	7.289
Cetane number improvers	Di-tert-butyl peroxide	(DTBP)	3.138	-3.376	36.23
	Methyl Oleate	(MO)	7.002	2.177	44.47
	Ethylene glycol	(EG)	6.497	3.154	-5.623
	Oleic acid	(OA)	12.47	3.134	31.82
	Stearic acid	(SA)	11.84	3.924	62.24

4.5.3 Characterization of the Additive Molecules

Infrared (IR) spectroscopy-based characterization was used to determine the chemical constituents or molecular structures of the additive training set that describe the orientation and alignment of these molecules. Using the NIST Webbook [39], the complete IR spectral region (4,000-400) cm^{-1} for twelve additives including antioxidants, cetane enhancers, and cold-flow improvers

were obtained. Ideally, both IR and NIR data would be available for the training set to provide better predictions, but unfortunately the availability of NIR data for the components of interest is severely limited.

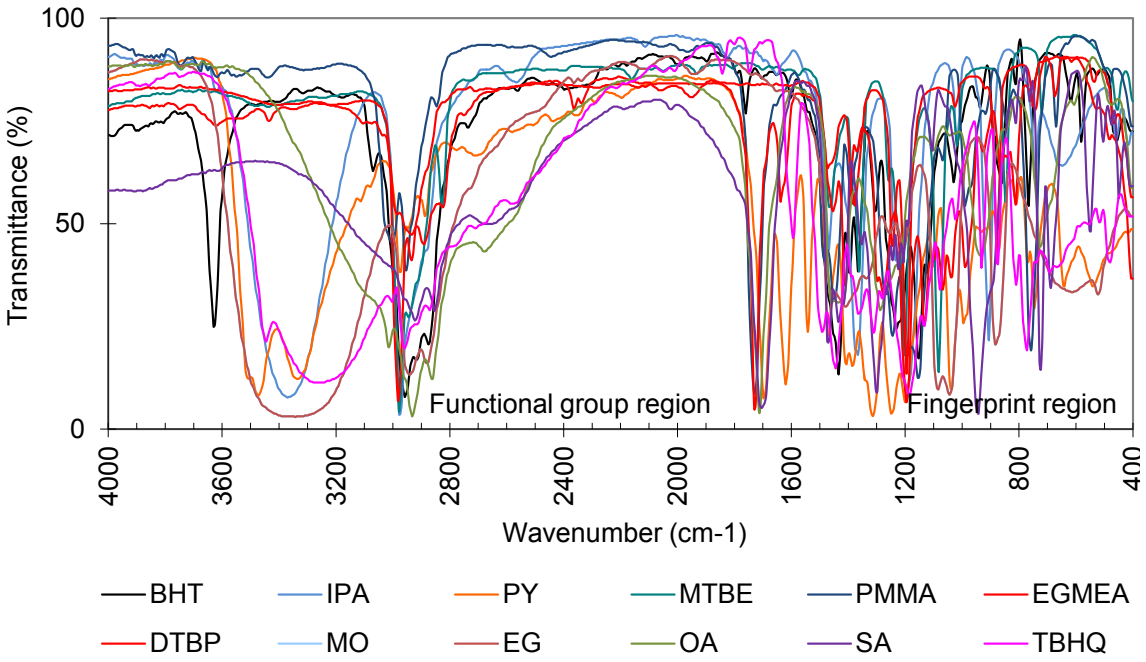


Figure 4.8: Infrared spectra of diesel additive molecules.

Figure 4.8 is IR spectra of the diesel additive molecules tabulated in Table 4.5 that serve as the descriptors of their molecular architecture. The IR spectra were translated to discrete variables by a process of *digitization*. In this process, each spectrum is fragmented into small equal fragments (2 cm^{-1} resolution) along the wavelength axis resulting with 1801 frequencies (descriptor variables).

4.5.4 Additive IR Data Analysis

The principal component scores were used to describe the variation in the multivariate characterization data with a minimum of variables to elucidate the underlying structure of the data. Principal components (PCs) captured the most variation possible in the smallest number of dimensions and

consolidate multiple property effects into single, underlying latent variables which are devoid of collinearity. The first three principal components captured 69.53% of the total variance of the standardized IR data (Figure 4.9).

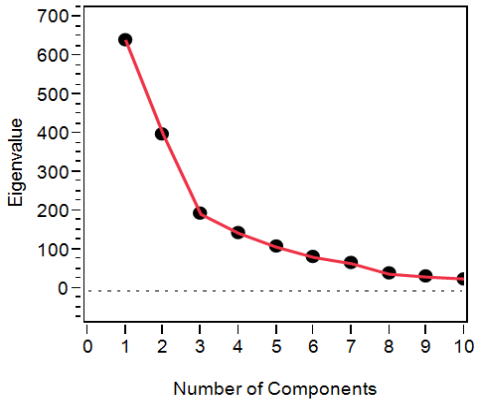


Figure 4.9: Scree plot for PCA on additive IR data.

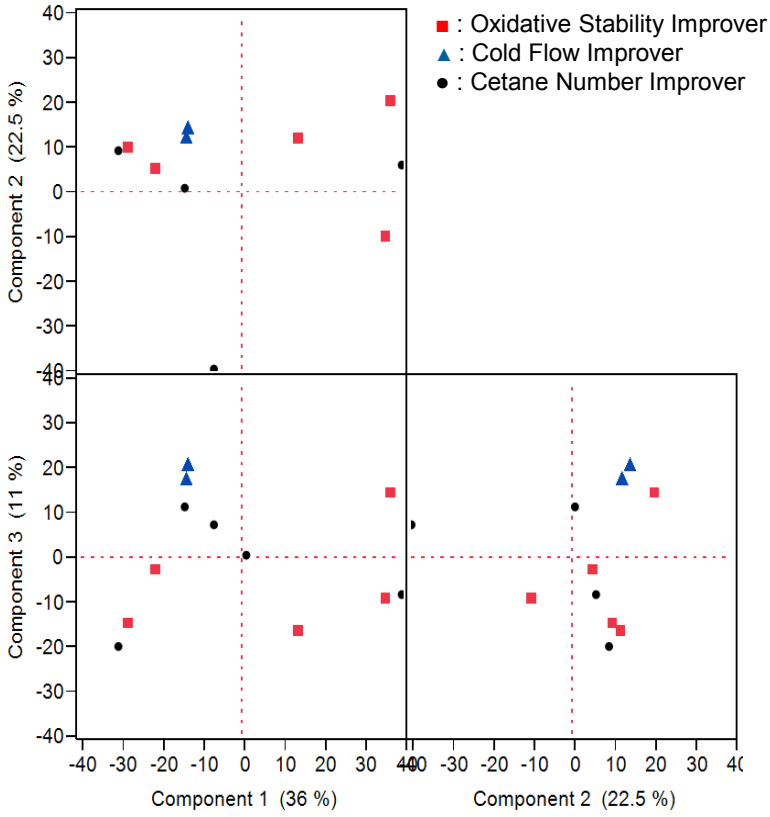


Figure 4.10: PCA score plots on first third PCs for additive IR data.

Figure 4.10 is a score plot that maps the three different additives types involved in the training set. Additives that are close to each other should have similar properties, whereas those far from each other should be dissimilar with respect to the three properties of interest. It is clear from the picture that cold flow additives (▲) are together in all the plots. However, oxidative stability (■) and cetane improvers (●) are scattered. This may suggest that an additive that improves oxidative stability also improves cetane and vice versa. This information is supported by the trend depicted in Figure 4.5.

4.5.5 Latent Variable Model Development

As each score vector t_i is a linear combination of the initial predictor variables **X** (the IR data), nonlinear regression models that describe response variables **Y** (the fuel properties) subjected to the **X** values are obtained. First, the PCA for variable reduction and second, multiple regression for calibration model development could be considered as a nonlinear PCR (NPCR). Nonlinear Principal Component Regression resulted in second order regression models for melting point (Eq. (4.7)) and kinematic viscosity (Eq. (4.8)), whereas, a third order regression model was developed for cetane number (Eq. (4.9)). The generalized forms of these latent variable property models are as follows:

$$\exp\left(\frac{T_m}{T_{mo}}\right) = \beta_0 + \sum_{i=1}^{3} \beta_i t_i + \sum_{i<j}^{3} \sum_{j>i}^{3} \beta_{ij} t_i t_j \qquad \textbf{(4.7)}$$

$$\ln(\eta) = \beta_0 + \sum_{i=1}^{3} \beta_i t_i + \sum_{i<j}^{3} \sum_{j>i}^{3} \beta_{ij} t_i t_j \qquad \textbf{(4.8)}$$

$$CN = \beta_0 + \sum_{i=1}^{3} \beta_i PC_i + \sum_{i=1}^{3} \beta_{ii} t_i^2 + \sum_{i<j}^{3} \sum_{j>i}^{3} \sum_{k>j}^{3} \beta_{ijk} t_i t_j t_k \qquad \textbf{(4.9)}$$

The left hand sides (LHS) of the above models are forced to have a particular form such that the respective properties follow linear additive

rules as described by the general group contribution model equation (Eq. (2.13)). Table 4.6 tabulates the regression coefficients for the latent variable models represented by Eq. (4.7) through Eq. (4.9). Figure 4.11 is a plot comparing the actual versus the model predicted properties.

Table 4.6: Model coefficients using PCR.

Properties	$exp(T_m/T_{mo})$	$ln(\eta)$	CN
β_0	8.1702	0.8132	65.89
β_1	0.0793	-0.0005	-0.2833
β_2	-0.1758	-0.0545	-0.0195
β_3	-0.0401	0.0021	-1.3906
β_{12}	0.0048	0.0033	-
β_{13}	0.0063	-0.0024	-
β_{23}	0.0144	-0.0015	-
β_{11}	-	-	-0.0491
β_{22}	-	-	0.3826
β_{33}	-	-	-0.0632
β_{123}	-	-	0.0038
R^2	0.816	0.764	0.790
R^2_{adj}	0.595	0.480	0.422

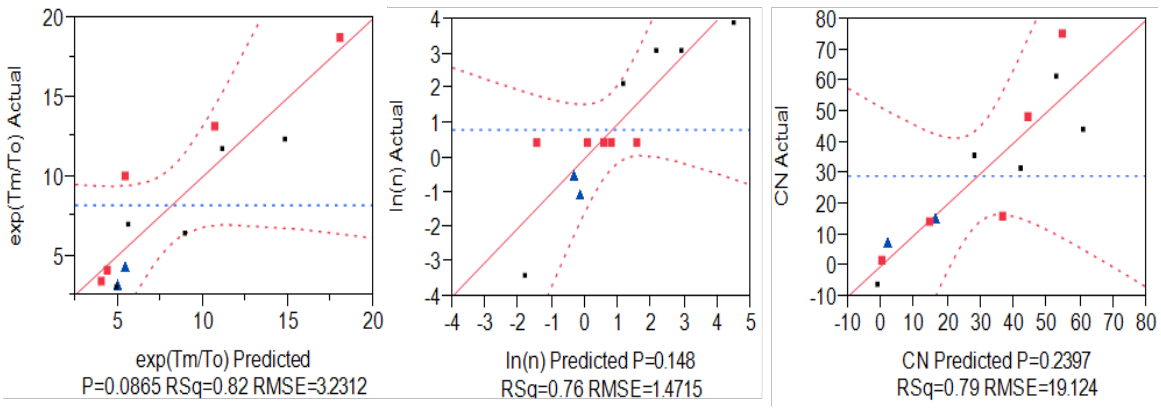


Figure 4.11: Predicted vs. actual product properties using PCR model.

4.5.6 Translating Physical properties to Latent Properties

The purpose of this case study is to identify candidate additives which, when mixed with a crude (off-spec) biodiesel, produce a fuel that meets the required fuel specifications presented in Table 4.7. Table 4.7 has much more stringent property constraint criteria than the one defined in Figure 4.7. Here, kinematic viscosity of 6.0mm^2/s at 40^{o}C is the maximum limit in ASTM D6751 (Table 4.4).

Table 4.7: Biodiesel target properties.

	Fuel Property		
	T_m [C]	η [mm^2/s] @ 40^{o}C	CN
Lower Limit	-60	4.51	47
Upper Limit	-20	6.00	65

In order to ensure that the target property space is properly explored, a feasibility region on the ternary diagram was evaluated. The identity of the exact shape of the feasibility region without extensive enumeration was evaluated by the six unique points described by Eq. (2.12). The three physical properties (p_i) were converted to latent properties (t_i) by solving Eq. (4.7), (4.8) and Eq. (4.9) simultaneously. The score variables are then standardized (q_i) using Eq, (2.43). The results are tabulated in Table 4.8. The feasibility region was mapped from the three-dimensional volume to a two-dimensional area utilizing the property clustering technique described in section 2.2.2. Figure 4.12 shows the biodiesel target feasibility region described by Table 4.8.

The minimum and maximum latent properties that correspond to physical properties in Table 4.7 are in Table 4.9.

Table 4.8: Physical and latent properties describing feasibility region.

Feasibility Region Points	Physical Property			Latent Property					
				Score			Standardized Score		
	T_m [K]	η [mm²/s]	CN	t_1	t_2	t_3	q_1	q_2	q_3
min,min,max	213	4.5	65	81	-21	-33	0.180	-0.048	-0.073
min,max,max	213	6.0	65	80	-21	-34	0.184	-0.049	-0.078
min,max,min	213	6.0	47	80	-20	-32	0.177	-0.043	-0.070
max,max,min	253	6.0	47	74	-19	-31	0.183	-0.048	-0.078
max,min,min	253	4.5	47	75	-20	-30	0.178	-0.046	-0.071
max,min,max	253	4.5	65	75	-21	-33	0.186	-0.052	-0.081

Table 4.9: Biodiesel target latent properties.

	Latent Property		
	q_1	q_2	q_3
Lower Limit	0.174	-0.054	-0.088
Upper Limit	0.191	-0.043	-0.065
Reference value	0.06	-0.02	-0.05

The properties of a crude biodiesel feedstock were selected such that it does not meet the fuel specification target properties defined in Table 4.9. The standardized latent property values are tabulated in Table 4.10. The crude biodiesel along with the petro-diesel property values are mapped onto the cluster space as shown in Figure 4.12.

Table 4.10: Crude biodiesel properties.

	Latent Property		
	q_1	q_2	q_3
Crude biodiesel	0.142	-0.0480	-0.0820

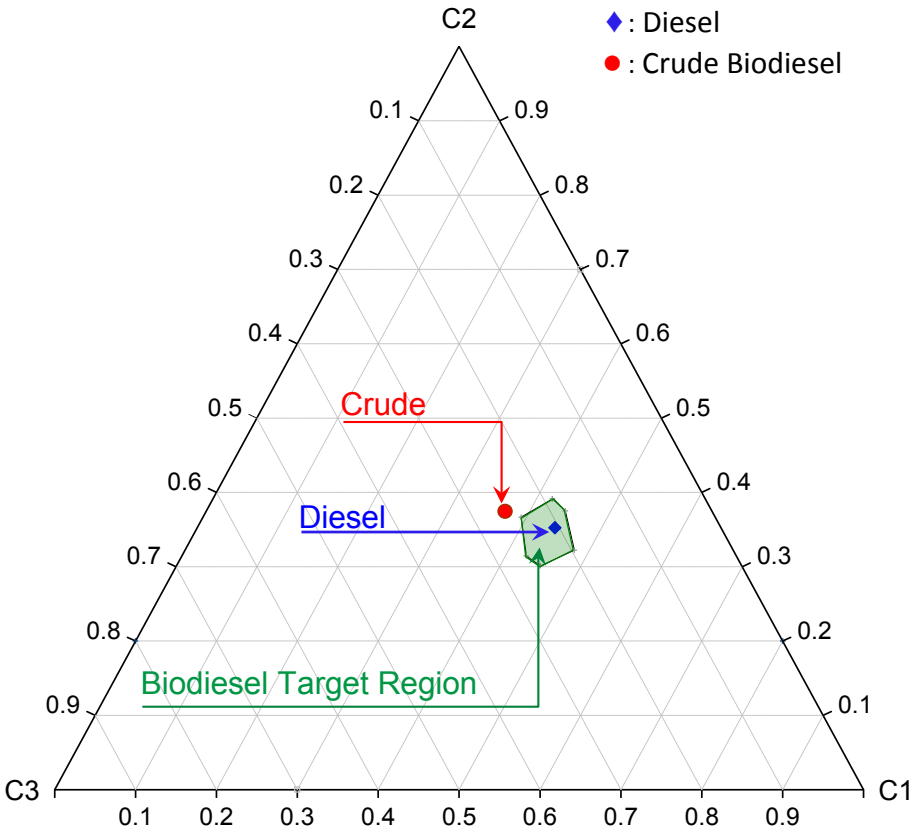


Figure 4.12: Target feasibility region and crude biodiesel in cluster space.

4.5.7 Evaluation of Desired Additive Feasibility Region

An infinite number of possible additives may exist that could be mixed with crude biodiesel. However, it is difficult to narrow down the feasible additive candidates such that the final mixture products meet the target biodiesel properties. Since the property cluster formulation enables linear mixing and lever arm analysis of the latent properties, it is straightforward to identify the feasibility region for the additives. The region bounded by the black dashed lines (shaded area) in Figure 4.13 represents the entire latent property search space for the potential additive molecules. The corresponding latent property ranges are presented in Table 4.11. The additive feasibility

region will serve as the property target region in the molecular design algorithm.

Table 4.11: Biodiesel additive latent property feasibility region.

	Latent Property		
	q_1	q_2	q_3
Lower Limit	0.519	-0.076	-0.138
Upper Limit	0.174	0.000	0.000

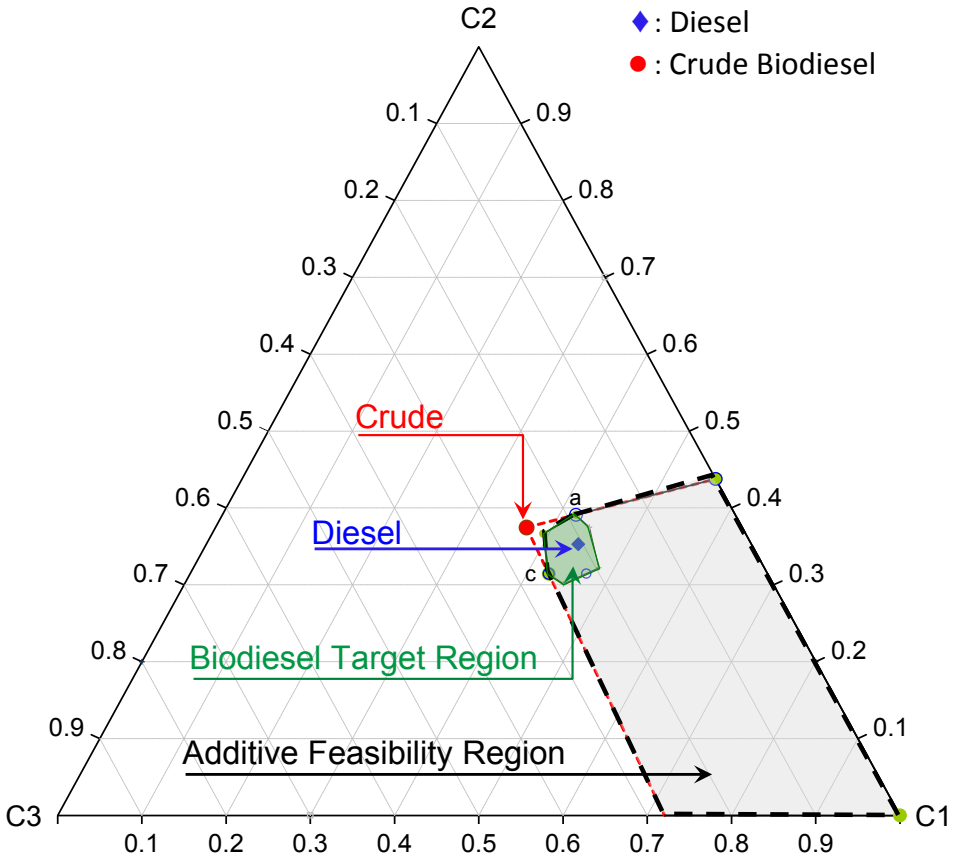


Figure 4.13: Desired additive design feasibility region in cluster space.

4.5.8 Enumeration of Desired Additive Molecules

The characterization-based group contribution method (cGCM) developed by Solvason [27] was used to take advantage of the additional molecular architecture information provided by characterization data. Characterization-based molecular groups/fragments are linearly combined to formulate a molecule. The property of the formulated molecule is determined by the individual contributions of each molecular group that make up the molecule. Twenty-three molecular groups were selected from the additive training set. Some of the groups are the fundamental building blocks present in every additive molecule and represent first order contributions, while the rest are larger groups and represents higher level of contributions. These selected molecular groups are considered to be a set of basic groups which represent the chemical make-up of the training set and are listed in Table 4.12 .

The infrared (IR) descriptor data [**X**-matrix] for the molecular fragments was compiled from Socrates [84] and is listed in Appendix B, Table B.2. The latent property contributions of each group are evaluated using the loading matrix ($\mathbf{P}^T$) obtained from the PCA of the training set data such that:

$$\mathbf{T}_{M\times A} = \mathbf{X}_{M\times K} \cdot \mathbf{P}_{K\times A} \tag{4.10}$$

The standardized score values (q_i) are calculated as described previously (Eq. (2.43)) and the results are tabulated in Table 4.12. Visual Basic for Application (VBA) codes are used to enumerate all potential molecules from the characterization-based groups in Table 4.12 that satisfied the target property constraints described in Table 4.11. Using the group-based property model, molecular groups or fragments are added together analogous to inter-stream conservation. A maximum number of similar groups, $N_g = 2$, was selected such that progressive combinations of similar groups are added until the maximum is reached.

Table 4.12: Molecular groups and their latent property contributions.

S.N.	Molecular Group	Latent Property		
		q_1	q_2	q_3
1	Methine Group, -CH- -	0.04	0.459	0.005
2	Methylene Group, -CH_2-	-0.05	-0.12	0.003
3	Methyl Group, -CH_3	-0.04	0.156	-0.025
4	Tetramethyl Group, -$C(CH_3)_3$	0.025	0.359	0.061
5	Aliphatic Methoxy Group, -O-CH_3	-0.03	-0.66	0.061
6	Vinyl Group, -CH=CH_2	0.125	-0.43	0.106
7	Vinylidene Group, CH_2=C- -	-0.04	0.339	0.002
8	cis-Vinylene Group, -CH=CH-	-0.18	-0.17	0.084
9	trans-Vinylene Group, -CH=CH-	-0.17	-0.18	0.091
10	Hydroxyl Group, -OH	-0.02	-0.37	-0.229
11	Primary Alcohol Group, -CH_2OH	0.093	-0.21	-0.086
12	Secondary Alcohol Group, - -CHOH	0.09	0.275	-0.056
13	Alilphatic Ether Group, -O-	0.15	-0.27	0.072
14	Alkyl Peroxide Group, -O-O-	0.211	-0.34	0.134
15	Saturated Aliphatic Ester Group, -CO-O-	0.247	0.328	0.155
16	Saturated Aliphatic Methyl Ester Group, -CO-O-CH_3	0.184	0.404	0.172
17	Saturated Aliphatic Ethyl Ester Group, -CO-O-CH_2CH_3	0.162	0.5	0.139
18	Acrylate Ester Group, CH2=CH-CO-O-	0.13	0.179	0.14
19	Methacrylate Ester Group, CH2=C(CH_3)-CO-O-	0.111	0.547	0.105
20	o-Alkyl Phenol Group (With H-bonding)	0.006	0.455	-0.042
21	p-Alkyl Phenol Group (With H-bonding)	0.007	0.42	-0.042
22	Monosubstituted Benzenes	-0.1	-0.27	0.081
23	1,2,4- Trisubstituted Benzene	0.055	-0.42	0.069

The enumerated candidate molecules were then screened for structural constraints, such as free bond number (*FBN*), to ensure that a stable, connected molecule was formed. The candidate molecular structures that were identified to fall within the feasibility region for the additive properties in Figure 4.14 and satisfy all the constraints, are presented in Table 4.13. The list of potential molecules is mapped onto the cluster space and is shown in Figure 4.14.

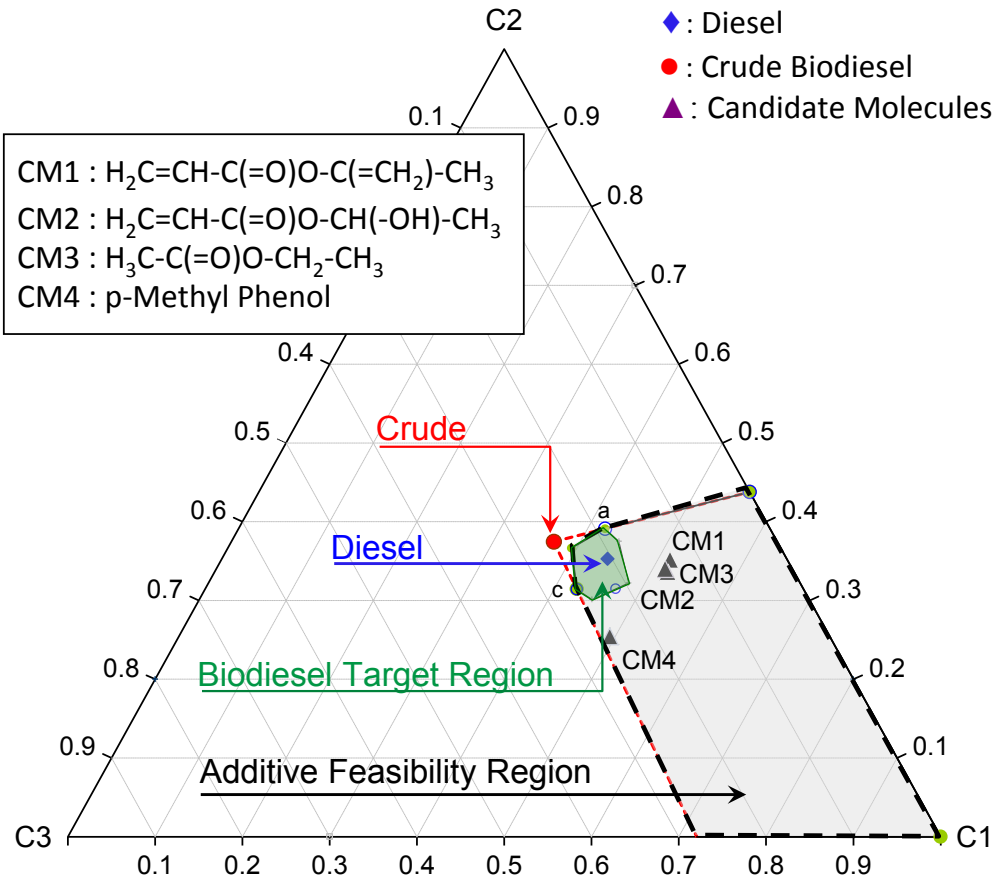
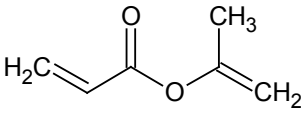
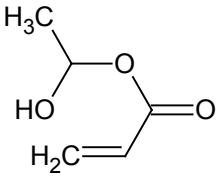
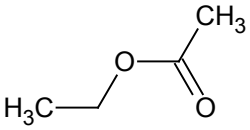
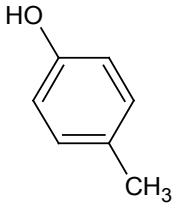


Figure 4.14: Cluster diagram for biodiesel blending problem.

Table 4.13: Results from characterization-based molecular design.

I.D.	Candidate Molecules	T_m [K]	η [mm²/s]	*CN*
CM1	 (Isopropenyl acrylate)	203	0.400	3.63
CM2	 (Hydroxyethyl acrylate)	261	1.04	3.04
CM3	 (Ethyl acetate)	172	0.72	2.92
CM4	 (4-methyl phenol)	306	9.68	7.07

In addition, solubility parameters can be used as a simple method to predict and calculate the dissolving power of the above candidate additive molecules in biodiesel as a further screening criterion. As a general rule, two substances with close solubility parameters (δ) should be mutually soluble [85]. In the thermodynamics of solution, the Hansen solubility parameters related to dispersion force (δ_d), polar interaction (δ_p), and hydrogen bond interaction (δ_h) have been conveniently used to estimate the solubility or miscibility between two compounds. These parameters can be estimated from additive group contributions [85]:

$$\delta_d^i = \frac{\sum F_{di}}{V_m} \qquad \delta_p^i = \frac{\sqrt{\sum F_{pi}^2}}{V_m} \qquad \delta_h^i = \sqrt{\frac{\sum E_{hi}}{V_m}} \tag{4.11}$$

where F_{di}, F_{pi}, and E_{hi} are contributions from group i for calculating dispersion, polar, and hydrogen component solubilities respectively using Hoftyzer and Van Krevelen method [85]. The molar volume (V_h) of a molecule was estimated by group contribution methods [34]. The total Hansen solubility parameter (equivalent to Hildebrand solubility parameter) can be expressed as:

$$\delta = \sqrt{\delta_d^2 + \delta_p^2 + \delta_h^2} \tag{4.12}$$

Table 4.14: Additive solubility in FAME at 25°C.

I.D.	δ_d MPa$^{1/2}$	δ_p MPa$^{1/2}$	δ_h MPa$^{1/2}$	δ MPa$^{1/2}$	V_m cm^3/mol	Feasible
CM1	13.75	3.59	7.16	15.91	136.68	Yes
CM2	12.44	5.12	14.05	19.45	107.35	No
CM3	10.97	3.59	7.16	13.58	97.80	Yes
CM4	19.55	4.77	13.78	24.39	105.37	No
FAME	15.7	1.46	4.57	16.43	335.36	-

Table 4.14 tabulates the solubility parameters and molar volumes of the additives and a common fatty acid methyl ester (FAME). Since linoleic acid (C18:2) represents a major constituent in the fatty acid profile (Table 4.2), it is used to check the solubility of the candidate additive molecules in methyl ester (biodiesel). Additives with lower a solubility parameter than FAME were considered miscible in most proportions with FAME ($\delta \leq 16$ MPa$^{1/2}$). Based on this criterion, the candidates CM1 and CM3 satisfied the screening criteria and were considered feasible solutions.

From Table 4.14 it can be inferred that among the Hansen parameters, the high dispersion parameter values (δ_d) have the major attractive factor for FAME and the additive molecules involved in the process of solubilization. These interactions arise from induced dipoles and their strength is related to the polarizabilities of the molecules (dipole moment of FAME, CM1, and CM3 molecules are primarily from the dipole moment of the carbonyl group). The polar parameter (δ_p) and the hydrogen bonding parameter (δ_h) of FAME, CM1, and CM3 are comparatively low.

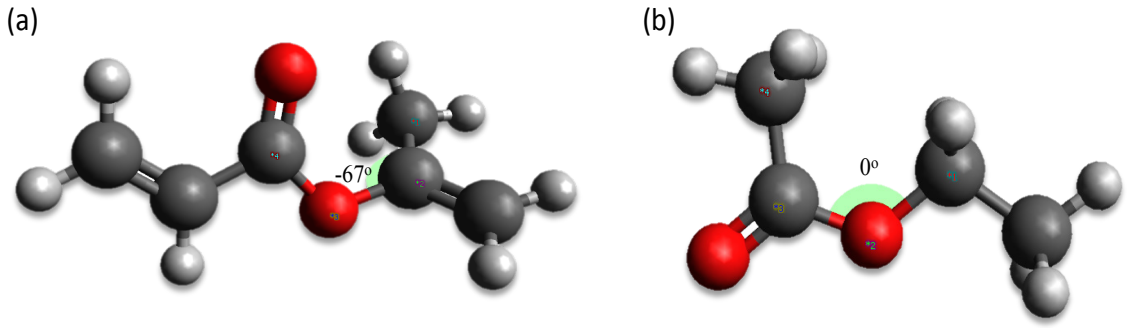


Figure 4.15: Spatial representation of candidate additive molecules according to B3LYP/6311++G(3df,3dp) calculations: (a) isopropenyl acrylate (CM1), and (b) ethyl acetate (CM3).

Furthermore, density functional theory (DFT) calculations were performed for the CM1 and CM3 additive molecules to optimize their geometry. Using the B3LYP method with the 6311++G(3df,3dp) basis set the lowest energy conformer for each molecule is presented in Figure 4.15. The estimated dipolar moments of CM1, CM3 and FAME are 1.54, 2.11 and 4.51 respectively. This is in agreement with the polar parameters obtained with Hansen's theory (Table 4.14). The difference between dipole moments ($|\mu_{FAME} - \mu_{Additive}|$) is at a minimum for CM1, which corresponds to the better affinity between CM1 and FAME and to the highest solubility limits of this additive. Consequently, we can conclude that both the polarity and spatial

configurations of feasible additive molecules are involved in the process of solubilization.

4.6 Conclusion

Unlike in group contribution methods (GCM), where parameters of the contribution are obtained by fitting the group contribution model to experimental data for a set of chemical compounds, characterization-based group contribution method (cGCM) will account for the chemical information thus extending the predictive capabilities of this method. Additionally, insights into the molecular structures of the candidates and candidate mixtures were obtained by incorporating the dipole moment (an important descriptor of a molecule) indirectly through the use of IR as a descriptor of molecules.

It should be noted that the limited amount of data and supporting characterization information available in the training set (and thus the molecular fragments available for molecular design) impacts the chemical stability/feasibility of the molecules that are generated in this step. Additional training set data can improve the quality of the predictions and thus increase the application range but would not require a different optimization methodology.

CHAPTER 5

REVERSE DESIGN OF IONIC LIQUIDS

5.1 Introduction

Design of environmentally benign solvents and alternative media for extraction and purification are new challenges within chemical product design. One class of novel compounds being studied for such application is ionic liquids (ILs) [11, 86, 87, 88]. They have become the subject of an increasing number of investigations due to their unique properties such as high polarity, stability at high temperature, flame resistance, and negligible vapor pressure. Ionic liquids that have tailored structures with an array of unique functional properties can have important applications in areas such as CO_2 capture and sequestration, sulfur removal from fuels, energy storage, biomass pre-treatment, and chemical separations. Through variation of both cation and anion, particular ionic liquids with tunable physical properties can be tailored. For example, the miscibility of ionic liquids with water or organic solvents can be varied with alkyl chain lengths on the cation and the type of anion present.

It is estimated that over 10^{14} unique cation/anion combinations are possible for use as room temperature ionic liquids, the majority of which have never been synthesized [87]. Thus, it is essential to develop a logical and systematic approach of selectively choosing a given ionic pair that matches a set of desired physico-chemical property targets. However, the traditional experimental trial-and-error approach of searching through this large molecular space is unrealistic as it is both time and labor intensive. An

example of possible combinations of cation, anion, and the alkyl chain of the side chain attached in the cation is represented graphically in Figure 5.1.

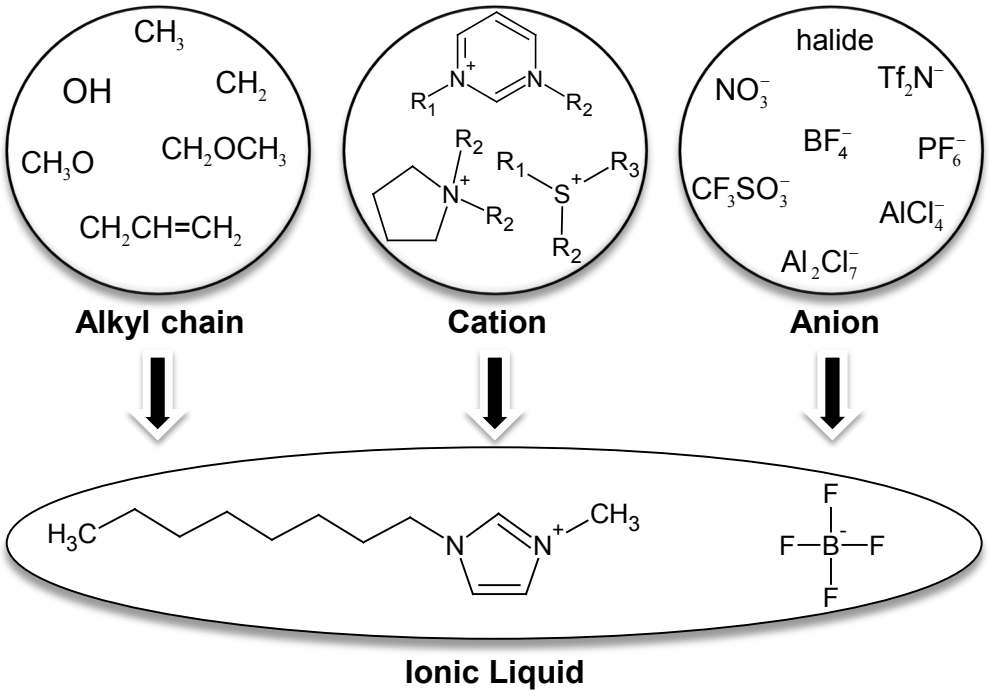


Figure 5.1: Selection of anions, cations, and side chains attached in cations for a task specific ionic liquid application.

Computer-assisted approaches towards the automated design of chemically formulated products with desired physical, chemical and biological properties hold immediate potential. In such methods, predicting the properties of ILs would be necessary for molecular design of such compounds.

The advantage of the newly developed characterization-based group contribution technique [27] has previously been demonstrated to predict physical-chemical properties and design of biodiesel additive molecules (Section 4.5.3) [53]. By exploiting the fact that molecules absorb specific frequencies that are characteristic of their structure, vibrational spectroscopy is used to elucidate chemical constituents, and the orientation and alignment of molecules. For example, infrared (IR) spectroscopy-based characterization

contains large quantities of descriptor data involving information on molecular architecture at atomic-, nano-, and micro-scales to describe physical properties and attributes of chemical products. Some common characterizations used to quantify molecular architectures at other length scales include nuclear magnetic resonance (NMR) and x-ray diffraction (XRD) (See Section 2.4 for more detail). In addition, the characterization-based group contribution method (cGCM) utilizes the latent property parameters based on characterization data instead of conventional regression-based property parameters which often exhibit poor attribute-property relationships [27].

The choice of an appropriate spectra training set is vital to building the latent variable structure since the training set defines the molecular architecture building blocks that can be used in the cGCM. In order for the method to be independent of the availability of experimental spectroscopic data, in this Chapter we investigated the use of density functional theory (DFT) based simulation techniques to generate the required IR spectra as molecular descriptors to develop predictive property models that can be used for the reverse design of ionic liquids.

5.2 Density Functional Theory

A set of ionic liquids with three properties were compiled from IUPAC Ionic Liquid database (ILThermo) [89] and listed in Table B.3. For each of these ionic liquid, the ion-paired structures were drawn into the Accelrys Draw 4.0 program to develop an initial 2-dimensional MDL Molfile. This information was then imported into Avogadro (v1.0.3) [90] to develop an initial geometry utilizing a molecular mechanics (MM) force-field for energy minimization. The MM force-field chosen for this task was MMFF94, also known as the Merck Molecular Force-Field. MMFF94 has been parameterized for a wide range of organic chemistry calculations and several charged molecules have

been included as well, most notably imidazolium cations [91]. For IL's with phosphate containing anions, the unified force-field (UFF) was utilized to generate starting point geometries because the MMFF94 force-field is not suited to handle these types of molecules.

With this starting point geometry, the final molecular geometry optimization was performed using the quantum chemical Gaussian 09 program [92] executed through a supercomputing cluster run by the Alabama Supercomputing Authority. As a molecular model to simulate the pure ionic liquid, ion-paired structures combining the cations and anions were optimized as a whole. The DFT method B3-LYP utilizing the 6-311G(2d,p) basis set in the ideal gas phase was used for molecular geometry optimization of all ion-paired structures in Table B.3. The method and basis set were chosen because similar computational level (B3-LYP/6-311+G(2d,p)) has previously been demonstrated to have the lowest root-mean-square error (when compared to Hartee-Fock (HF) method utilizing similar basis sets) for simulated fundamental molecular vibrations [93]. Figure 5.2 shows the optimized structure of the stable gas phase (local potential energy minimum) 1-ethyl-3-methyl imidazolium hexafluorophosphate [emIm]PF_6 ionic liquid conformer.

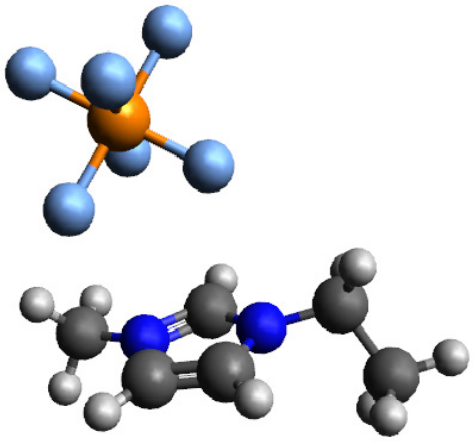


Figure 5.2: Geometry optimized molecular structure of [emIm]PF_6 at the B3LYP/6-311G(2d,p) computational level.

Vibrational frequency calculations were performed for each ion-paired structure to confirm the presence of an energy minimum. This is confirmed as the resulting frequencies were all positive values (no imaginary frequencies) [93]. Negative values appearing in this data would reveal the presence of a transition state geometry since the vibrational modes are derived from a square root of the force constants from the Hessian matrix (Eq. B.6).

In general, the values of harmonic vibrational frequencies determined by *ab initio* computational methods are larger than the experimental frequencies because the methods neglect anharmonicity effects, incompletely incorporate electron correlations and uses finite basis sets [94]. The vibrational spectra calculated using DFT (which incorporates electronic correlation) requires a correlation factor of 0.965 at the B3-LYP/6-311G(2d,p) level [95]. The base set is a mathematical function to approximate the electronic wave function. Figure 5.3 (a) shows the high correlation (R^2 of 0.998) of the IR frequency calculated to the available experimental data [93] for [emIm]PF_6. Figure 5.3 (b) shows the resulting simulated vibrational spectra using Gaussview software.

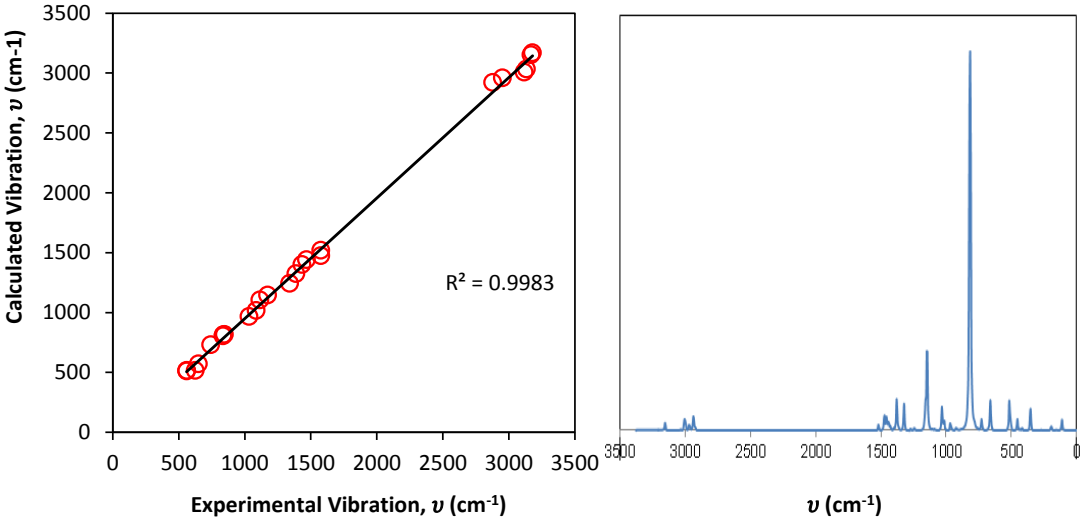


Figure 5.3: (a) Calculated vs. experimental IR frequencies and (b) infrared spectrum at 2cm^{-1} resolution of [emIm]PF_6 at the B3-LYP/6-311G(2d,p) computational level.

Table 5.1 contains the theoretical and experimental vibrations along with vibrational assignments for [emIm]PF_6.

Table 5.1: B3LYP/6-311+G(2d,p) vibrational assignments (cm^{-1}) of [emIm]PF_6.

	Frequency (cm^{-1})	
Assignment	**Calculated**	**Experimental**
Cat-An tors	21	-
Cat-An bend	39	-
Cat-An bend	56	-
Cat-An bend	62	-
Cat-An bend, CH_2(N) bend	99	-
at-An bend, CH_2(N) bend	106	-
CH_3(N) twist	127	-
CH_3(N) twist	162	-
CH_2(N) bend	186	-
terminal CH_3 twist	210	-
CH_3(N) bend	218	-
PF_6 scissors	284	-
PF_6 scissors	287	-
PF_6 scissors	295	-
CCH bend, CH_3(N) bend	296	-
CCH bend, CH_3(N) bend	326	-
CCH bend, CH_3(N) bend	389	-
PF_6 scissors	434	-
PF_6 scissors	439	-
PF_6 scissors	440	-
FPF sym str	447	-
FPF sym str	472	-
FPF sym bend	527	-
FPF sym bend	528	-
FPF sym bend	531	559
ring ip sym bend	569	559
ring op asym bend	570	625

Assignment	Frequency (cm^{-1})	
	Calculated	**Experimental**
ring op asym bend	593	649
FPF sym str, ring op bend	695	-
ring ip bend, CH_3(N) bend, CH_2(N) bend	710	-
ring HCCH sym bend	749	741
CCH bend	812	-
ring HCCH asym bend	826	836
FPF asym str, ring HCCH asym bend	875	838
FPF asym str, ring NC(H)N bend, CCH bend	900	847
FPF asym str	905	-
ring NC(H)N bend	907	-
CC str	960	-
ring ip sym str	981	-
ring sym str, CH_3(N) str, CH_2(N) str	1012	1033
CC str, ring ip sym str	1063	1087
ring HCCH sym bend, ring ip sym str	1082	-
ring HCCH sym bend, ring ip sym str	1138	-
CC str	1159	1114
CH_3(N) HCH bend	1173	-
ring sym str, CH_3(N) str, CH_2(N) str	1200	1172
ring ip asym str, CC bend	1216	-
ring ip asym str, CH_3(N) str	1267	-
ring ip sym str, CH_2(N) str	1298	1340
ring ip sym str, CH_2(N) str, CH_3(N) str	1328	-
ring ip asym str, CH_2(N) bend	1382	1387
CC str	1429	-
ring ip asym str	1447	-
ring ip asym str, CH_3(N) str	1468	1432
CCH HCH sym bend, CH_3(N)HCH sym bend	1505	-
CCH HCH asym bend, CH_3(N)HCH sym bend	1513	1468
CH_3(N) asym bend	1522	-
CC HCH bend	1533	-
CH_3(N) HCH sym bend	1538	-
ring ip sym str, CH_3(N) str, CH_2(N) str	1547	1575

Assignment	Frequency (cm^{-1}) Calculated	Experimental
ring ip asym str, CH_3(N) str, CH_2(N) str	1591	1575
terminal CH_3HCH sym str	3059	2878
CH_3(N) HCH sym str	3065	-
CH_2 HCH sym str	3076	-
terminal CH_3HCH asym str	3102	2952
CH_2 HCH asym str	3109	-
CH_3(N) HCH asym str	3137	-
CC HCH asym str	3145	-
CH_3(N)HCH asym str	3157	3115
ring NC(H)NCH str	3176	3134
ring HCCH asym str	3300	3168
ring HCCH sym str, ring NC(H)NCH str	3316	3179

5.3 Data Analysis and Model Development

IR data of 22 training set IL molecules generated using DFT contained 701 descriptor variables at 2 cm^{-1} resolution for (3500-0) cm^{-1} frequency range. Using PCA, the first three principal components captured about 60% of the total variance. The score values are listed in Table 5.2. The loading values are not presented as they constitute a matrix of [701x3]. Using PCR, the latent variable property models (Figure 5.4) were developed. The generalized expressions are shown in Eq. 5, 6, and 7 for dynamic viscosity, density, and melting temperature, respectively. Table 5.3 contains the respective model coefficients. Note that the log, inverse, and exponent transformations of the calibrated QSPR models were made to ensure linear mixing of the property operators of these properties.

$$\ln(\mu) = \beta_0 + \sum_{i=1}^{3} \beta_i t_i + \sum_{i=1}^{3} \beta_{ii} t_i^2 \qquad \textbf{(5.1)}$$

$$\frac{1}{\rho} = \beta_0 + \sum_{i=1}^{3} \beta_i t_i + \sum_{i=1}^{3} \beta_{ii} t_i^2 \qquad \textbf{(5.2)}$$

$$\exp\left(\frac{T_m}{T_{mo}}\right) = \beta_0 + \sum_{i=1}^{3} \beta_i t_i + \sum_{i<j}^{3} \sum_{j>i}^{3} \beta_{ij} t_i t_j + \sum_{i=1}^{3} \beta_{ii} t_i^2 \qquad \textbf{(5.3)}$$

Table 5.2: PCA score values for X- and Y-blocks.

X score value		
t_1	t_2	t_3
-3.102	-1.676	2.218
37.75	-0.137	6.502
25.57	-1.434	5.686
-21.98	-10.25	7.347
-10.86	-1.346	-11.18
-0.315	-1.392	3.867
-0.923	1.408	-8.155
28.34	1.497	3.089
6.778	0.023	-7.867
-0.438	-0.708	1.950
18.46	9.313	-5.078
-6.854	-1.547	-7.027
-6.392	0.682	-11.33
-5.741	-0.417	-6.179
18.04	-1.227	4.456
-16.84	-11.40	8.355
-13.73	-3.260	2.640
-23.92	-5.838	9.794
-36.60	-1.853	7.743
-8.058	0.288	-10.72
-25.84	29.45	10.24
14.50	-1.365	3.655
-4.452	-0.656	-2.259

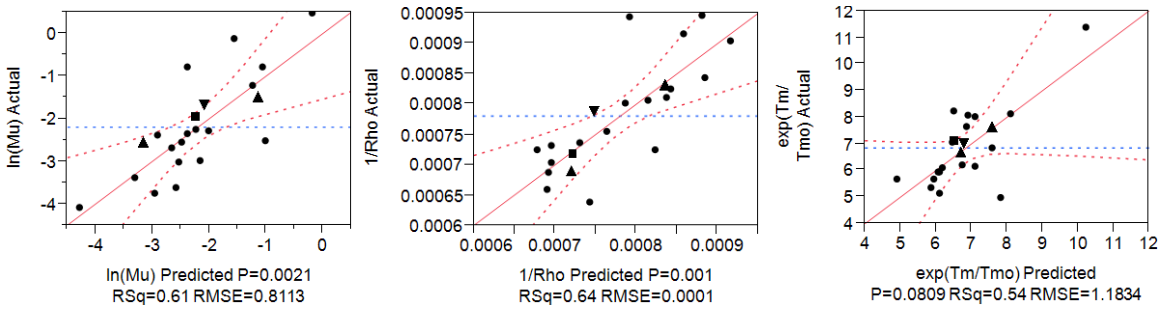


Figure 5.4: Predicted vs. actual IL properties using PCR model.

Table 5.3: Model coefficients using PCR.

Properties	$ln(\mu)$	$1/\rho$	$\exp(T_m/T_{mo})$
β_0	-1.273	8.13e^{-4}	6.778
β_1	-0.0736	-6.03e^{-7}	-0.0311
β_2	0.315	-3.25e^{-6}	0.0152
β_3	0.175	1.05e^{-5}	-0.0206
β_{13}	-	-	-0.0137
β_{23}	-	-	0.0293
β_{11}	-	-1.14e^{-7}	0.0035
β_{22}	-0.0162	-	-0.0168
R^2	0.609	0.643	0.541
R^2_{adj}	0.517	0.559	0.312

The poor accuracy of predictions for the properties of interest could be related to the quality of the experimental data used for calibrating the property models. Moreover, the gas phase nature of IR simulation has difficulties to take into account the structural features of ionic liquids in liquid state (electrostatic, van der Waals interactions, hydrogen bonds, etc.). In addition, accounting for the impact of variation in the structure of ions (size, symmetry, conformational flexibility, etc.) on properties is always challenging [31]. For example, melting point (T_m) of ILs decreases with increase in the length of alkyl substituents due to reduction of electrostatic

interactions between ions. However, this trend is not always true. With the length of alkyl group increasing, van der Waals attraction between bulk alkyl radicals favors an increase in (T_m). Experimental observations have shown oscillation of the melting point of ILs with the size of alkyl groups [96]. In addition, characterization of ionic liquids using the vibrational spectroscopy techniques like IR works better on molecules with covalent bonds. Although, ionic liquids have some degree of covalent bonding, the ionic character dominates the covalent character.

5.4 Reverse Design of Ionic Liquids using QSPR and cGCM

Identification of ionic liquids that possess task specific properties through time consuming experiments and simulations of individual alternative molecules is virtually impossible. However, there are relatively small numbers of distinct building-blocks/functional-groups/fragments that can theoretically be combined to generate a wide variation of possible molecules. In order to capture the group contribution and interaction variability in the 22 training ILs (listed in Table B.3), a total of fifteen groups are selected. Table 5.5, Table 5.6, and Table 5.7 contain seven anion, six cation, and two alkyl chain cation substituents, respectively. These selected molecular groups are considered to be a set of basic groups which represent the chemical make-up of the training set.

Anion groups are considered as a whole molecule without any possible attachment position. Cation groups consist of imidazolium, pyridinium, and alkyaminium bases with a maximum of four possible alkyl group attachments. Two alkyl groups are selected as possible alkyl chains attached to cation groups as these were the only groups present in the training set. All these groups are considered first order in the group contribution-based property estimation of the formulated ionic liquid molecule. Since most of the groups cover a significant proportion of a potential IL molecule, they

incorporate most of the interactions between the groups. Moreover, the use of projection methods for property model calibration also capture the most common features and underlying latent relationships resulting from group-group, ion-ion, and ion-group interactions among the ionic pairs in the training set molecules. The loadings (**P**) from PCA preserve the covariance structure which can be applied to reverse design of ionic liquid molecules that are consistent with structural attributes in the training set molecules.

The infrared frequency data for the anion, cation, and alkyl groups are generated using B3-LYP/6-311G(2d,p) computational level and by following the procedure outlined in Section 5.2. These IR descriptors values are then transformed to score values using $\mathbf{T} = \mathbf{X} \cdot \mathbf{P}$. Finally, the three latent properties are obtained by normalizing the score values to follow linear mixing in terms of property operators (see Section 2.7 for detail). The latent property values which are the contributions of each group to the three properties of interest (μ, ρ, and T_m) are tabulated in Table 5.5, Table 5.6, and Table 5.7.

As a proof-of-concept example, an ionic liquid is to be designed as an alternative benign solvent that should replace a molecular solvent with environmental and health hazard properties. Table 5.4 presents the target property ranges in terms of viscosity, density, and normal melting temperature for a certain task-specific application. The corresponding latent property values were obtained following the steps outlined in Section 2.6.2.

Table 5.4: Target ionic liquid properties.

	Physical Properties			**Latent Properties**		
	μ [Pa.s]	ρ [kg/m^3]	T_m [K]	q_1	q_2	q_3
Lower Limit	1.0	1200	230	0.0004	-0.0077	-0.0346
Upper Limit	1.1	1300	300	0.0671	0.4807	0.0248

Table 5.5: Anion groups and their latent property contributions.

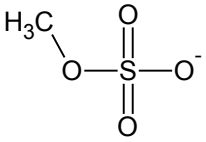
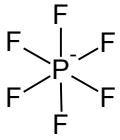
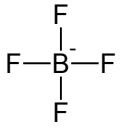
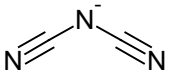
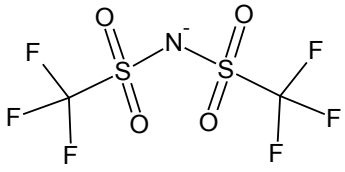
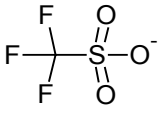
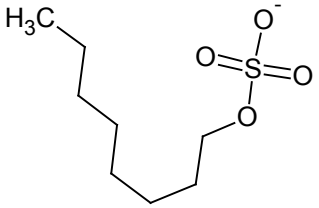
Anion Groups	Latent Property		
	q_1	q_2	q_3
 Methylsulfate	0.43	2.45	-0.08
 Hexafluorophosphate	0.15	-1.91	-0.14
 Tetrafluoroborate	0.14	-2.92	0.74
 Dicyanamide	0.27	7.70	-0.71
 Bis(trifluromethylsulfonly)-amide	0.11	3.51	1.40
 Triflate	-0.06	-2.67	0.70
 Octylsulfate	-0.04	-5.17	-0.90

Table 5.6: Cation groups and their latent property contributions.

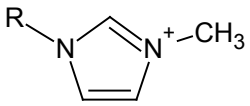
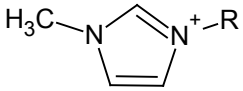
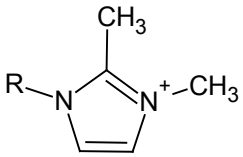
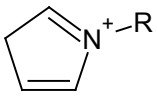
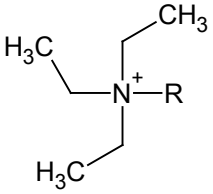
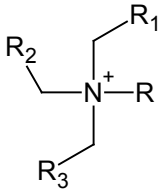
Cation Groups	Valency	Latent Property		
		q_1	q_2	q_3
 methyl imidazolium	1	-0.06	0.73	0.16
 methyl imidazolium (+)	1	-0.20	-0.27	0.00
 2,3 dimethyl imidazolium	1	0.55	0.12	0.37
 Pyridinium	1	-0.43	0.33	-0.16
 Triethy alkyaminium	1	0.91	0.99	0.67
 Trialkyl alkyaminium	4	0.23	-0.90	-0.04

Table 5.7: Alkyl chain attached to cation groups and their latent property contributions.

Alkyl Groups	Valency	Latent Property		
		q_1	q_2	q_3
—CH_3 Methyl	1	-3.08	0.08	-0.37
–CH_2 – Methylene	2	4.08	0.92	1.37

The reverse design of potential IL molecules is accomplished by exhaustively generating combinatorial structures from the given molecular fragments until the resulting properties match the target property values in Table 5.4. In exhaustive searches, selection from among numerous permutations of anion, cation, and alkyl chain attached to cation groups is performed. First, an anion is selected, and the properties of interest are evaluated by using a group-based property estimation method, changing the cations and the length of side alkyl chain attached to the cation. For every cation; CH_3 groups, equal to the free bond number (*FBN*) of the cation, are added as a cap at the free end of the cation. Finally, varying numbers of CH_2 groups are added until the sum of property values of all the groups fall within the target property range. In this case study, a maximum of fifteen CH_2 groups are allowed to occur in a generated ionic liquid pair. Figure 5.5 depicts a scenario for the reverse design of ILs. The molecular property is estimated based on the first order group contribution method:

$$\Omega_j = \sum_{g=1}^{15} n_g \cdot \Omega_g \qquad \textbf{(5.4)}$$

where Ω_j is the normalized latent property operator of property j (which includes 3 properties), n_g is the similar number of group appearances, and Ω_i is the normalized latent property contribution of the appeared group.

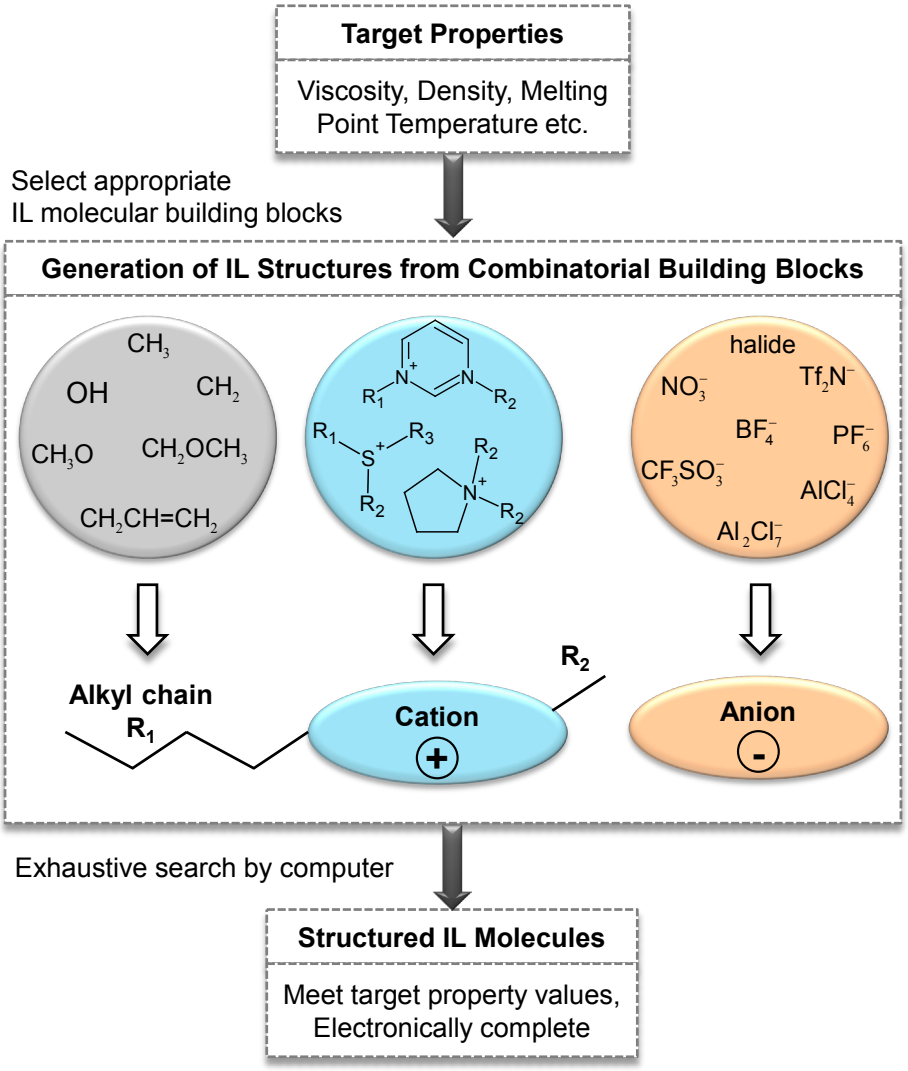


Figure 5.5: Scenario of reverse design of ILs.

A heuristic and exhaustive search algorithm in the Python programming language was written to automate the generate-and-test procedure to enumerate all possible IL molecules by selecting collections of groups, estimating their physical properties, and checking these against the target values. A total of 10,638 possible candidate were generated. Only 26 of these enumerated IL molecules satisfied the target property specifications in latent space (Table 5.4) and the free bond number (*FBN*) structural constraint. The structures whose *FBN* equals to zero ensure feasible, stable, and connected

molecules are formed (see Section 2.8.2). Out of 26 of these molecules, 23 solutions were not present in the training data set (Table B.3). Table B.4 contains 23 unique solutions in latent or principal property space. When these 23 IL molecules were mapped from principal property space to physical property space, only 13 IL molecules satisfied the target physical properties (Table 5.4). It must be noted that all 23 solutions are feasible; however, only 13 are feasible in both the spaces. The loss of 10 molecules when properties of the IL molecules were transformed from the latent space to physical space could be contributed to the uncertainties associated with the property models.

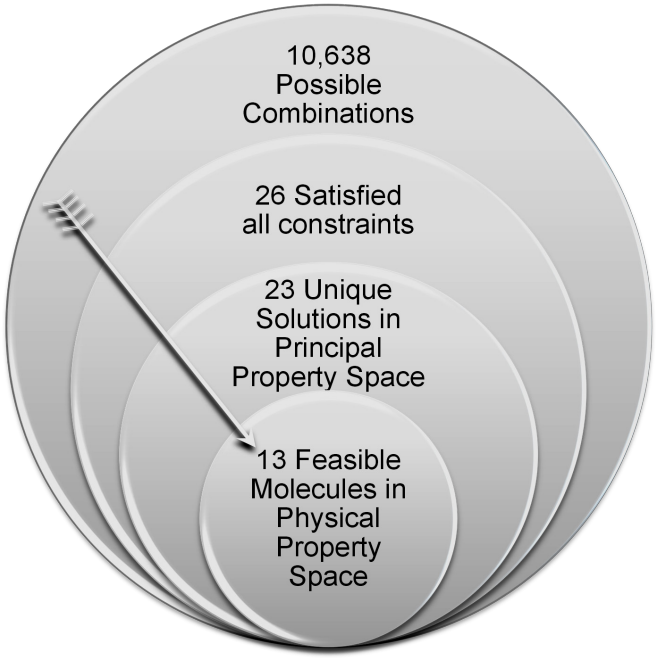


Figure 5.6: Enumerated and validated candidate IL molecules.

Table 5.8 lists the final candidate IL molecules for which predicted properties matches the target properties and are structurally feasible. The ID numbers and candidate molecule names are the same as in Table B.4. These molecules could further be subjected to experimental and/or more detailed computational simulation tests. The enumeration and selection after validation of IL molecules is schematically presented in Figure 5.6.

Table 5.8: Thirteen candidate ionic liquid molecules that match target properties in property space.

ID	Candidate Molecule	Physical Properties		
		μ [Pa.s]	ρ [kg/m³]	T_m [K]
3	$[e_3pNH_4]mSO_4$	1.00	1248	267
4	$[meIm]PF_6$	1.00	1226	290
7	$[meIm]BF_4$	1.00	1261	270
8	$[e2,3m_2Im]BF_4$	1.00	1261	266
9	$[ePy]BF_4$	1.00	1291	260
11	$[e_3bNH_4]BF_4$	1.00	1247	267
12	$[e_3pNH_4]dCN$	1.00	1259	261
14	$[e2,3m_2Im]Tf_2N$	1.00	1247	267
17	$[e_3bNH_4]Tf_2N$	1.00	1233	247
18	$[meIm]CF_3SO_3$	1.00	1263	266
21	$[e_3bNH_4]CF_3SO_3$	1.00	1221	277
22	$[ePy]oSO_4$	1.10	1203	252
23	$[e_3pNH_4]oSO_4$	1.00	1255	262

5.5 Conclusion

Ionic liquids (ILs) as green solvents can be used in separation processes, chemical synthesis, catalysis and electrochemistry, successfully replacing the conventional volatile, flammable and toxic organic solvents. Within the computer-aided molecular design (CAMD) framework, a characterization based method was combined with chemometric techniques towards the design of IL structures corresponding to particular physical properties. Infrared spectra (IR) generated from density functional theory (DFT) simulations were used as molecular descriptors for capturing information on molecular architecture, and for calibration of latent variable property models to design ILs in a logical and systematic methodology. The use of DFT eliminates the dependency on the availability of experimental IR data,

thereby extending the capabilities of a design method based on such characterization techniques.

In addition, the design of ILs using the characterization-based group contribution method (cGCM) further demonstrated the advantages of using it compared to the conventional GCM for predicting properties. For new class of chemical compounds such as ILs, tabulated group contribution values for many molecular groups are not available in the literature. Here, cGCM expands the application range of general GCM to predict properties of the enumerated IL molecules from the combination of IL molecular fragments by using the molecular information captured from characterization technique based on infrared spectroscopy data.

CHAPTER 6

FUTURE WORK

6.1 Methodology Improvements

The chemical product formulation methodology presented in this dissertation utilizes multivariate data analytics in three areas: (1) characterization, (2) modeling, and (3) design. Each of these aspects in product design plays an important role at their respective levels. The contribution of this work has been in the development of a generalized methodology that integrates all of the above tools and techniques to design new and improved materials with tailored properties. The applications are presented in the form of case studies involving the design of biofuel additives, thermoplastic formulation, and ionic liquid. Several of the tools and the techniques combined in this work may not be an ideal choice; however, the general problem formulation as well as the solution approach can be extended to other problems in chemical engineering.

In the future, some of the areas of improvement can be seen in each of the multivariate analytics. Figure 6.1 presents schematics of the potential tools and techniques in the characterization, modeling, and design.

6.1.1 Multi-Dimensional Characterization

In silico molecular design approaches presented in the foregoing chapters can take advantage of an increasing amount of techniques available to characterize molecular architectures. Recently, the advancement of high performance computers and robust theories, such as quantum chemistry, information theory, graph theory, etc., are accelerating a paradigm shift in

the molecular modeling. Today, thousands of molecular descriptors that capture and transform the information encoded in the molecular structure are effortlessly generated using readily available algorithms and software. Each molecular descriptor takes into account a small part of the whole chemical information contained within a real molecule. It is important to consider the whole environment of the compound as a potential source of information to investigate its interplay with physicochemical properties and biological activities. QSPR/QSAR modeling is an integral part of *in silico* molecular/drug design approaches.

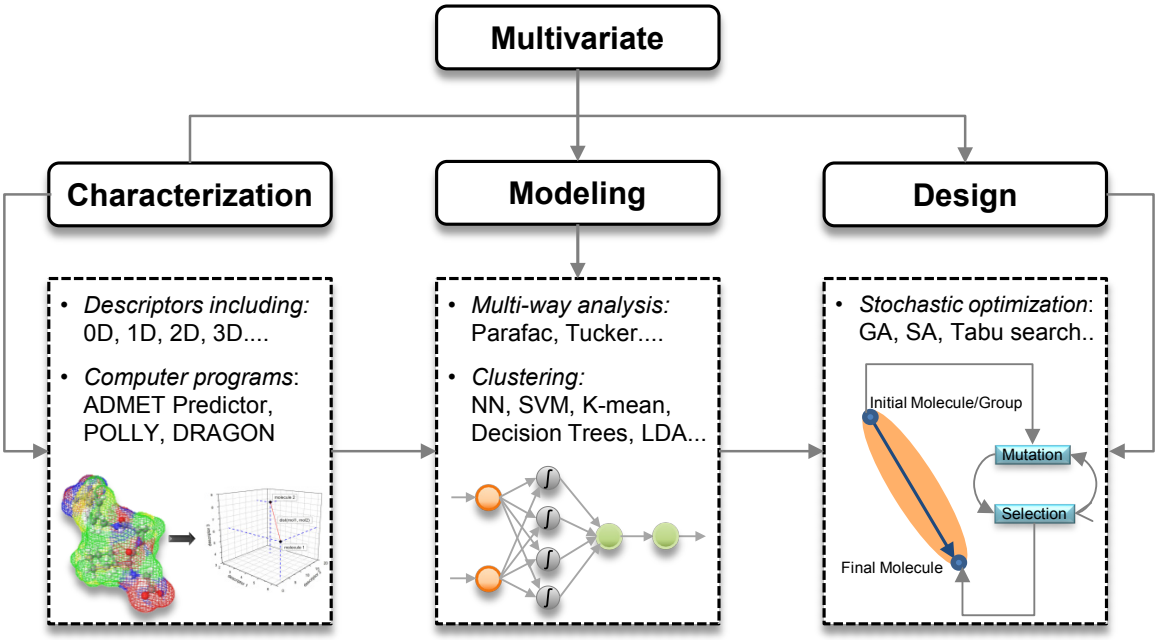


Figure 6.1: Improvements in the process of multivariate characterization, modeling, and design.

Since knowledge gained from chemical data alone is inadequate for success in drug discovery, it is important to closely combine or merge chemoinformatics and bioinformatics in the future works. Among many computer programs, DRAGON™ computer software can generate more than 1,600 molecular descriptors, covering topological, molecular, and 3D

properties of molecules [97]. E-DRAGON applet is an electronic version of DRAGON and is freely available at www.vcclab.org/edragon/. Likewise, ADMET Predictor™ is a powerful software that can estimate a number of properties associated with physicochemical and biological properties of drug-like chemicals and their molecular structures [98]. The addition of descriptors with diverse chemical features and wider molecular space coverage would make the presented methodology more flexible and robust, but would not require a new problem formulation and solution framework.

In addition, since spectroscopic methods are fast and non-invasive, their use has grown considerably in recent decades in both research and industries such as food, petrochemical and pharmaceutics production. To establish more chemical features of samples, two distinct analytical methods can also be combined. For example, chromatography can be applied first to separate the components, while infrared spectroscopy can be applied later to identify them. When two distinct methods are combined, in so-called *hyphenated methods*, the resulting data is multivariate and multi-way. Likewise, fluorescence excitation-emission matrix (EEM) spectroscopy is a flexible, rapid, and portable organic matter characterization tool. In order to use these characterization techniques, however, two-way data decomposition techniques such as principal component analysis (PCA) and partial least square (PLS), used in this dissertation, cannot be applied without further extension.

6.1.2 Multi-Way Modeling

Decomposition of multi-way data arrays can be accomplished by using *parallel factor analysis* (PARAFAC), which is an extension of the two-way PCA to three-way systems. Unlike the bi-linear PCA model, the tri-linear PARAFRAC model factors do not have rotational freedom, and thus the solution is essentially unique [99, 100]. Figure 6.2 (a) and Figure 6.2 (b) show a schematic of two-way data and three-way tensor decomposition,

respectively. While PCA has a score and loading matrix (Figure 6.2 (a)), PARAFAC has a score matrix and two loading matrices (Figure 6.2 (b)) for three-way systems.

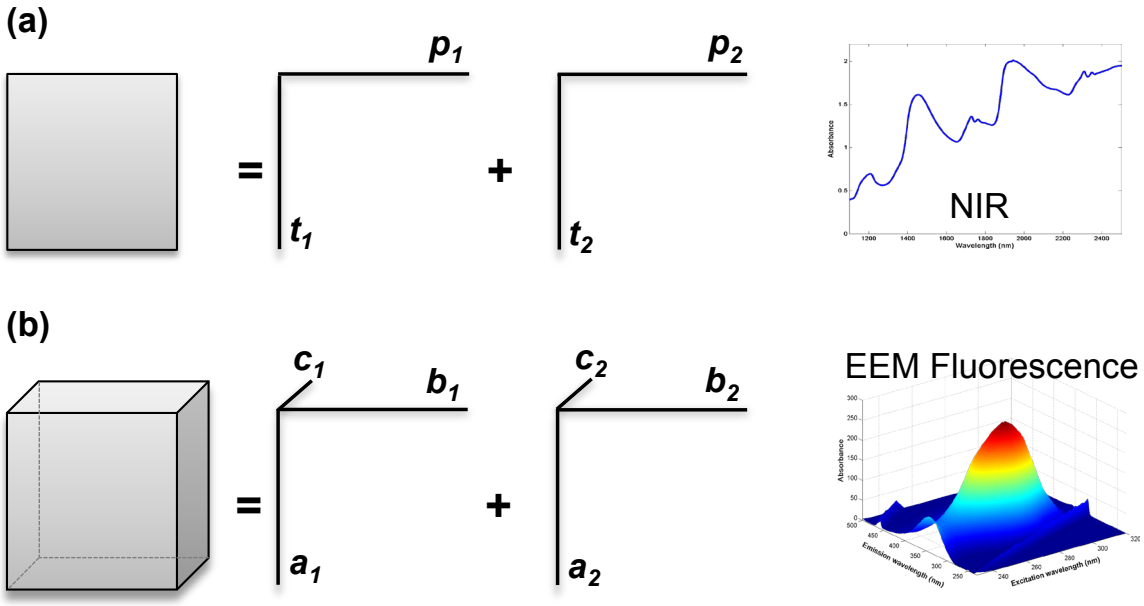


Figure 6.2: The decomposition of X-block by (a) PCA and (b) PARAFAC.

Another important area where PARAFAC becomes instrumental is in the batch-wise manufacturing processes that are common in chemical, pharmaceutical, and semi-conductor industries. Batch processes give rise to three-dimensional matrices as shown in Figure 6.2 (b). The mixture formulation problem presented in Chapter 3 involves a two-way data structure. The PARAFAC decomposition of a three-way tensor $\underline{\mathbf{X}}$ is represented by Eq. (6.1).

$$x_{i,j,k} = \sum_{a=1}^{A} a_{i,a} \cdot b_{j,a} \cdot c_{k,a} + e_{i,j,k} \qquad i = 1,....,I;\ j = 1,...,J;\ k = 1,...,K \tag{6.1}$$

For batch processes, Eq. (6.1) contains I batches, J time points, and K variables. For fluorescence EEM, Eq. (6.1) contains I samples with J emission wavelength and K excitation wavelengths. Analogous to PCR regression,

PARAFAC can be combined with MLR to obtain a multi-way regression model: decompose **X** using PARAFAC and regress **Y** on the PARAFAC scores **A**` as presented in Sections 2.6.2 and 2.6.3.

In the other hand, when linear approximations (such as PCA or PARAFAC) are not valid, neural networks (NNs), which use a series of weights (w_i) and hidden neurons to detect complex and non-linear relationships between inputs and outputs, can be utilized [101]. Figure 6.3 is a schematic for an artificial neural network. NNs are non-linear statistical data modeling tools suitable for high-dimensional and non-linear data such as the data generated from EEM fluorescence spectroscopy. They are an adaptive system that changes its structure based on external or internal information that flows through the network. They are useful when predictive accuracy is the most important objective.

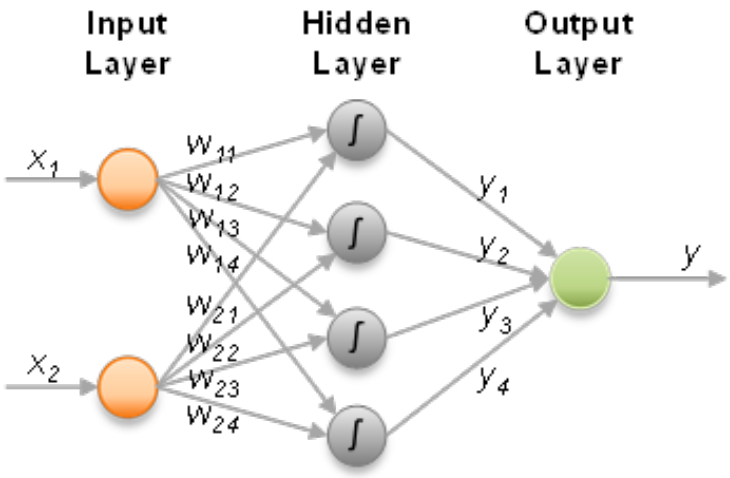


Figure 6.3: Structure of an artificial neural network.

6.1.3 Stochastic Search and Optimization

The deterministic solver is used in the molecular design problems presented in this dissertation, involving the design of biodiesel additives and ionic liquids. The developed methodology shifted the product design paradigm from

guess-and-test to generate-and-test approach. Molecules with desired properties are generated using dynamic programming that searches through molecular space defined by a set of combinatorial building blocks. The combinatorial explosion is minimized by using meta-groups (see Section 2.8.2). In group-based property estimation methods, these meta-groups are treated as first order groups. The problem formulation using a deterministic approach is able to handle the presented design problem because the molecular search space is small. However, when the molecular search space becomes combinatorially large and nonlinear, stochastic optimization algorithms like evolutionary algorithm [9], simulated annealing (SA), tabu search [11], and artificial neural networks (ANN) are more appropriate and effective computer algorithms.

Genetic algorithms (GAs) are stochastic evolutionary searches that use an analogy to chromosome encoding and Darwinian model of natural selection and evolution [9]. GAs are useful for combinatorial problems dealing with highly complex and a highly dimensional search space. Unlike traditional search methods that move in the objective area deterministically (point-by-point forward movement), GAs move probabilistically (parallel movement) in the optimal direction. In GAs, by stochastically favoring the mating of a more fit population of molecules, the most promising areas of the search space are explored at the expense of low performance regions [9].

6.1.4 Managing and Handling Uncertainty

Uncertainty accumulates through multiple steps in variable transformation and can make the variance of the final response undesirably large. The identification, quantification, and communication of model prediction uncertainties are important steps in determining the usefulness of any model. Uncertainty analysis techniques such as uncertainty propagation (forward and inverse) need to be incorporated into model calibration and design methodologies. By developing new design methods that are capable of

dealing with uncertainty in multi-scale models of materials and its propagation through subsequent design and analysis, the application range increases and design becomes robust to uncertainties in the design process. Importance of model uncertainty becomes greater as the extrapolation becomes farther from the historical or training data set.

6.2 Design of Inherently Benign Chemical Process Routes

Traditionally, process design was guided by two major facets, technical and economic decision criteria. However, the incorporation of two further dimensions of sustainability, the ecological and the social aspects, simultaneously into the early process design stages has become necessary in recent years [102]. Figure 6.4 depicts different aspects involved in early process design. Opportunities for identification and development of inherently safer process alternatives for solvents, reaction paths, catalysts, etc., are most abundant during the early design stage. An inherently safer process avoids or reduces hazards instead of controlling them [103, 104].

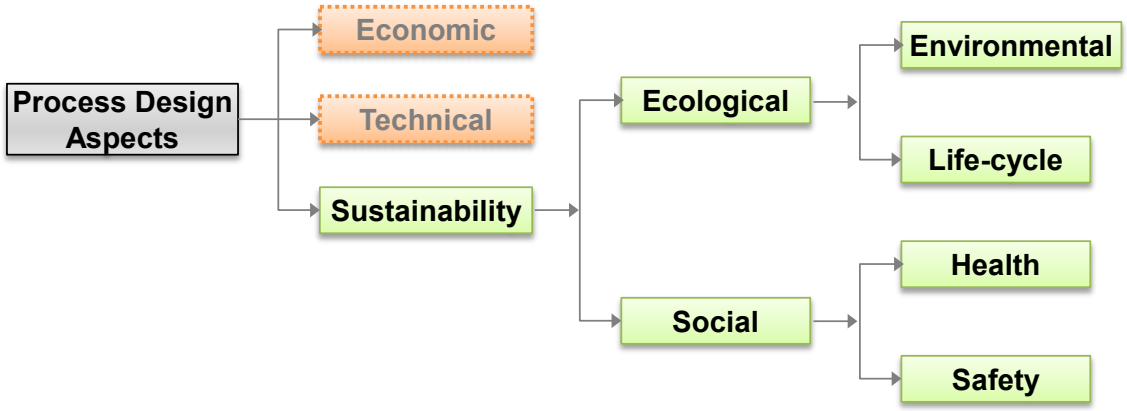


Figure 6.4: Aspects of early process design.

Analogous to the molecular architecture descriptors used in molecular design work, various process route descriptors related to environment, health, and safety factors such as toxicity, reactivity, flammability, heat of

reaction, temperature, process yield, properties of chemicals involved, etc., can be explored and used to design inherently safer synthesis pathways. Here, the primary objective is to minimize the amount of raw material usage, energy usage, waste generation, and the hazard to both life and environment.

Many index-based approaches exist to compare and rank routes based on environment, health, and safety (EHS) hazards and life cycle assessment (LCA) impacts, however, they generally suffer from scaling and multivariate related issues. The recently proposed statistical analysis based Inherent Benign-ness Indicator (IBI) framework by Srinivasan and Nhan [103] alleviates the above shortcomings. They considered principal components (PCs) as statistically independent factors of the routes instead of traditional empirical weighting schemes. The scores and loadings on the first two PCs make up the IBI and help characterize the inherent benign-ness of a process. The larger the IBI value, the less benign the route is.

The work in this dissertation can be easily extended to use this PCA-based IBI methodology to not only investigate the most inherently safer process route among different routes examined, but also to identify the similarities and differences in the EHS footprints of the available routes. This information and insight can be used to determine modifications needed to improve a route's benign-ness through molecular/mixture design. Reverse problem formulation could be applied as an integration procedure where the process unit performance that minimizes the IBI could be identified in the first step and then the molecule/mixtures that match the targets process performance would be identified in the second step.

The computer-aided molecular/mixture design (CAMD) framework combined with the property clustering technique would then be employed to generate products that meet the design constraints. If successful, the method could provide a new generation of greener chemical substitutes (e.g. solvents, catalyst, lubricants, heat transfer fluids, refrigerants, surfactants, etc.) that result in lower safety, health and environmental impacts. The algorithm will

be computationally efficient as it will be based on the targeting approach mentioned before in this dissertation. Moreover, simultaneously integrating inherent safety principles (using safer materials and operating conditions, minimizing inventory, etc.) into process and product design, from conception to completion, could lead to a more sustainable process.

REFERENCES

[1] Gani, R. (2004). Chemical product design: challenges and opportunities. *Computers & Chemical Engineering*, *28*(12), 2441-2457.

[2] Cussler, E. L., & Moggridge, G. D. (2001). *Chemical Product Design*. USA: Cambridge University Press, New York, ISBN 0521791839.

[3] Hill, M. (2009). Chemical Product Engineering - The third paradigm. *Computers & Chemical Engineering,* 33(5): 947.

[4] Ertl, P. (2003). Cheminformatics analysis of organic substituents: identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *Journal of Chemical Information and Computer Sciences, 43*(2), 374-380.

[5] Achenie, L. E. K., Gani, R., & Venkatasubramanian, V. (Eds.). (2002). *Computer Aided Molecular Design: Theory and Practice* (p. 408). Elsevier Science.

[6] Sahinidis, N. V., Tawarmalani, M., & Yu, M. (2003). Design of alternative refrigerants via global optimization. *AIChE Journal*, *49*(7), 1761–1775.

[7] Biegler, L. T., Grossmann, I. E., & Westerberg, A. W. (1997). *Systematic methods of chemical process design. New Jersey* (p. 796). Prentice Hall PTR.

[8] Camarda, K. V., & Maranas, C. D. (1999). Optimization in Polymer Design Using Connectivity Indices. *Industrial & Engineering Chemistry Research*, 38(10), 1884-1892.

[9] Venkatasubramanian, V., Chan, K., & Caruthers, J. M. (1995). Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm. *Journal of Chemical Information and Modeling*, *35*(2), 188–195.

[10] Harper, P. M., & Gani, R. (2000). Computer aided tools for design/selection of environmentally friendly substances. *Process Design Tools for Environment* (pp. 371–404). Philadelphia, USA: Taylor & Francis.

[11] McLeese, S. E., Eslick, J. C., Hoffmann, N. J., Scurto, A. M., & Camarda, K. V. (2010). Design of ionic liquids via computational molecular design. *Computers & Chemical Engineering*, *34*(9), 1476-1480. Elsevier Ltd.

[12] Contantinou, L., Gani, R. (1994). New group contribution method for estimating properties of pure compounds. *AIChE Journal*, 40(10): 1697-1710.

[13] Harper, P. M., & Gani, R. (1999). Computer-aided molecular design with combined molecular modeling and group contribution. *Fluid Phase Equilibria*, 337-347.

[14] Joback, K. G., Reid, R. C. (1983). Estimation of Pure-Component Properties from Group Contributions. *Chemical Engineering Communication* 57: 233.

[15] Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikstrom, C., & Wold, S. (2006). *Multi- and Megavariate Data Analysis: Basic Principles and Applications (Part I)* (2nd ed., p. 425). Umetrics Academy, Umea.

[16] Lavine, B. K. (2005). *Chemometrics And Chemoinformatics* (p. 216). American Chemical Society.

[17] Carrera, G., & Aires-de-Sousa, J. (2005). Estimation of melting points of pyridinium bromide ionic liquids with decision trees and neural networks. *Green Chemistry*, *7*(1), 20.

[18] Venkatasubramanian, V. (2009). Drowning in data: Informatics and modelling challenges in a data rich networked world. *AIChE Journal,* 55(1): 2-8.

[19] Wold, S. (1995). Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, *30*(1), 109-115.

[20] Jaeckle, C. M., & Macgregor, J. F. (1998). Product design through multivariate statistical analysis of process data. *AIChE Journal*, 44(5), 1105-1118.

[21] Qin, X., F. Gabriel, D. Harell, and M. M. El-Halwagi. 2004. Algebraic Techniques for Property Integration via Componentless Design, *Industrial & Engineering Chemistry Research,* 43 (14) (July): 3792–3798.

[22] Eden, M. R. (2003). *Property Based Process and Product Synthesis and Design*. (Doctoral dissertation). CAPEC, Department of Chemical Engineering, Technical University of Denmark, Retrieved from http://www.orbit.dtu.dk/fedora/objects/orbit85827/datastreams/file_5486802/content

[23] Shelley, M. D., El-Halwagi, M.M. (2000). Component-less design of recovery and allocation systems: a functionality-based clustering approach. *Computers & Chemical Engineering*, *24*(9-10), 2081-2091.

[24] Eljack, F. T., Solvason, C. C., Chemmangattuvalappil, N., & Eden, M. R. (2008). A Property-Based Approach for Simultaneous Process and Molecular Design. *Chinese Journal of Chemical Engineering*, *16*(3), 424–434.

[25] Chemmangattuvalappil, N. G., Eljack, F. T., Solvason, C. C., & Eden, M. R. (2009). A novel algorithm for molecular synthesis using enhanced property operators. *Computers & Chemical Engineering*, *33*(3), 636–643.

[26] Vlachos, D. G. (2012). Multiscale Modeling for Emergent Behavior, Complexity, and Combinatorial Explosion. *AIChE Journal,* 58(5), 1314-1325.

[27] Solvason, C. C. (2011). *Integrated Multiscale Product Design using Property Clustering and Decomposition Techniques in a Reverse Problem Formulation.* (Doctoral dissertation). Department of Chemical Engineering, Auburn University, Auburn, Alabama.

[28] Poling, B. E, Prausnitz, J. P., O'Connell, J. P. (2000). *The Properties of Gases and Liquids*, 5th edition, McGraw-Hill, New York, USA.

[29] Matsuda, H., Yamamoto, H., Kurihara, K., & Tochigi, K. (2007). Computer-aided reverse design for ionic liquids by QSPR using descriptors of group contribution type for ionic conductivities and viscosities. *Fluid Phase Equilibria*, *261*(1-2), 434–443.

[30] Linusson, a, Elofsson, M., Andersson, I. E., & Dahlgren, M. K. (2010). Statistical molecular design of balanced compound libraries for QSAR modeling. *Current medicinal chemistry*, *17*(19), 2001–16.

[31] Varnek, A., Kireeva, N., Tetko, I. V, Baskin, I. I., & Solov'ev, V. P. (2007). Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points? *Journal of chemical information and modeling*, *47*(3), 1111–22.

[32] Michielan, L., & Moro, S. (2010). Pharmaceutical perspectives of nonlinear QSAR strategies. *Journal of chemical information and modeling*, 50(6), 961–78.

[33] Linusson, a, Gottfries, J., Lindgren, F., & Wold, S. (2000). Statistical molecular design of building blocks for combinatorial chemistry. *Journal of medicinal chemistry*, *43*(7), 1320–8.

[34] Marrero, J., Gani, R. (2001). Group-contribution-based estimation of pure component properties. *Fluid Phase Equilibria*, 183-184: 183-208.

[35] Gabrielsson, J., Lindberg, N.-O., & Lundstedt, T. (2002). Multivariate methods in pharmaceutical applications. *Journal of Chemometrics*, *16*(3), 141-160.

[36] Gabrielsson, J., Lindberg, N.-O., Pålsson, M., Nicklasson, F., Sjöström, M., & Lundstedt, T. (2004). Multivariate methods in the development of a new tablet formulation: optimization and validation. *Drug development and industrial pharmacy*, *30*(10), 1037-49.

[37] Gabrielsson, J., Sjöström, M., Lindberg, N.-O., Pihl, A.-C., & Lundstedt, T. (2006). Multivariate methods in the development of a new tablet formulation: excipient mixtures and principal properties. *Drug development and industrial pharmacy*, *32*(1), 7-20.

[38] Workman, J., & Weyer, L. (2008). *Practical Guide to Interpretive Near-Infrared Spectroscopy* (1st ed., p. 344). CRC Press.

[39] S.E. Stein (2010), NIST Mass Spec Data Center, *NIST Standard Reference Database Number* 69, Eds. P.J. Linstrom and W.G. Mallard, National Institute of Standards and Technology.

[40] Eriksson, L., & Johansson, E. (1996). Multivariate design and modeling in QSAR. *Chemometrics and Intelligent Laboratory Systems*, *34*(1), 1–19.

[41] Kramer, R. (1998). *Chemometric Techniques for Quantitative Analysis.* New York (p. 203). CRC.

[42] Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185, 1–17.

[43] Kresta, J. V., MacGregor, J. F. and Marlin, T. E. (1991). Multivariate statistical monitoring of process performance. *Canadian Journal of Chemical Engineering*. 69, 35.

[44] Wold, S., Kettaneh, N., & Tjessem, K. (1996). Hierarchical multiblock PLS and PC models for Easier Model interpretation and as an alternative to variable selection. *Journal of Chemometrics*, 10, 463-482.

[45] MacGregor, J. F., and Kourti, T. (1995). Statistical process control of multivariate process. *Control Engineering Practice*. *3*(3), 403-414.

[46] Johnson, R. A. (2007). *Applied Multivariate Statistical Analysis* (6th ed., p. 800). Pearson.

[47] Jolliffe, I.T. (2002). Principal component analysis. New York: Springer.

[48] Jackson, J. E. (1991). *A User's Guide to Principal Components. Journal of the Operational Research Society* (Vol. 43, pp. 641-641). Wiley-Interscience.

[49] Jaeckle, C. M., & Macgregor, J. F. (1998). Product design through multivariate statistical analysis of process data. *AIChE Journal*, *44*(5), 1105-1118.

[50] Wentzell, P. D., & Vega Montoto, L. (2003). Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemometrics and Intelligent Laboratory Systems*, *65*(2), 257-279.

[51] Montgomery, D. C. (2000). *Design and Analysis of Experiments*, *5th Edition. America* (pp. 1-5). Wiley.

[52] Ben-Israel, A., & Greville, T. N. E. (2003). *Generalized Inverses: Theory and Applications*. (J. Borwein & P. Borwein, Eds.) *Book* (Vol. 18, p. 371). Springer.

[53] Hada, S., Solvason, C. C., & Eden, M. R. (2011). Molecular Design of Biofuel Additives for Optimization of Fuel Characteristics. In A. C. Pistikopoulos, E.N.; Georgiadis, M.C.; Kokossis (Ed.), *21nd European Symposium on Computer Aided Process Engineering* (pp. 1633–1637).

[54] Brignole, E., Bottini, S., & Gani, R. (1986). A strategy for design and selection for separation processes. *Fluid Phase Equilibria*, *13*, 331–340.

[55] Joback, K. G. (1989). *Designing Molecules Possessing Desired Physical Property Values*. (Doctoral dissertation). Massachusetts Institute of Technology. Retrieved from http://dspace.mit.edu/handle/1721.1/14191

[56] Kettaneh-Wold, N. (1992). Analysis of mixture data with partial least squares. *Chemometrics and Intelligent Laboratory Systems*, 14(1-3), 57-69.

[57] Muteki, K., & Macgregor, J. (2007). Multi-block PLS modeling for L-shape data structures with applications to mixture modeling. *Chemometrics and Intelligent Laboratory Systems*, *85*(2), 186-194.

[58] Cornell, J. (1990), *Experiments with mixtures*, 2nd Ed., Wiley, New York.

[59] Scheffe, H. (1963). Simplex-centroid design for experiments with mixtures. *Journal of the Royal Statistical Society*, 25 (2), 235-263.

[60] Cox, D. R. (1971). A note on polynomial response functions for mixtures. *Biometrica*, 58 (1), 155-159.

[61] Solvason, C. C., Chemmangattuvalappil, N. G., Eljack, F. T., & Eden, M. R. (2009). Efficient Visual Mixture Design of Experiments using Property Clustering Techniques. *Industrial & Engineering Chemistry Research*, *48*(4), 2245-2256.

[62] Muteki, K., MacGregor, J. F., & Ueda, T. (2006). Rapid Development of New Polymer Blends: The Optimal Selection of Materials and Blend Ratios. *Industrial & Engineering Chemistry Research*, *45*(13), 4653-4660.

[63] Muteki, K. (2006). *Mixture product design using latent variable methods*. (Doctoral dissertation). Department of Chemical Engineering, McMaster University.

[64] Grassmann, P., Sawistowski, H., & Hardbottle, R. (1971). *Physical principles of chemical engineering*. New York: Pergamon Press.

[65] JMP, Version 9.0. SAS Institute Inc., Cary, NC, 1989-2012.

[66] Janaun, J., & Ellis, N. (2010). Perspectives on biodiesel as a sustainable fuel. *Renewable and Sustainable Energy Reviews*, 14(4), 1312-1320. Elsevier Ltd.

[67] Schober, S., Mitelbach, M. (2004). The impact of antioxidants on biodiesel oxidation stability. *European Journal of Lipid Science and Technology*, 106, 382-389.

[68] Karmee, S.K., Chadha, A. (2005). Preparation of biodiesel from crude oil of Pongamia pinnata. *Bioresource Technology*, 96 (13), 1425-1429.

[69] American Society for Testing and Materials (ASTM), D6751-12 Standard Specification for Biodiesel Fuel Blend Stock (B100) for Middle Distillate Fuels, *ASTM International*, West Conshohocken, PA, 2012, www.astm.org.

[70] Fukud, H., Kondo, A., Noda, H. (2001). Biodiesel Fuel Production by Transesterification of Oils. *Journal of Bioscience and Bioengineering*, *92*(5), 405-416.

[71] Knothe, G., Gerpen, J. H. V., & Krahl, J. (2005). *The Biodiesel Handbook*. (G. Knothe, J. H. Van Gerpen, & Jurgen Krahl, Eds.) *Applied Sciences* (Vol. 2, p. 302). AOCS Press.

[72] Bröckel, U., Meier, W., & Wagner, G. (2007). *Product Design and Engineering: Best Practices*. John Wiley & Sons.

[73] Canakci, M., & Sanli, H. (2008). Biodiesel production from various feedstocks and their effects on the fuel properties. *Journal of industrial microbiology biotechnology*, 35(5), 431-441.

[74] Demirbas, A. (2003). Chemical and Fuel Properties of Seventeen Vegetable Oils. *Energy Sources*, 25(7), 721-728.

[75] European Committee for Standarization (CEN), EN 14214 (2003). Automotive fuels – fatty acid methyl esters (FAME) for diesel engines – Requirements and test methods, CEN, Brussels.

[76] Wang, P.S., Tat, M.E., Van Gerpen, J. (2005). The production of fatty acid isopropyl ester and their use as a diesel engine fuel. *Journal of the American Oil Chemists Society,* 82(11), 845-849.

[77] Lee, I., Johnson, L.A., Hammond, E.G. (1995). Use of branched-chain esters to reduce the crystallization temperature of biodiesel. *Journal of the American Oil Chemists Society*, 72, 1155-1160.

[78] Knothe. G. (2008), "Designer" Biodiesel: Optimizing Fatty Ester Composition to Improve Fuel Properties, *Energy and Fuels*, 22(2), 1358-1364.

[79] Ribeiro, M., Pinto, A. C., Quintella, C. M., Rocha, G. O., Teixeira, L. S. G., Guarieiro, L. N., Rangel, C., et al. (2007). The Role of Additives for Diesel and Diesel Blended (Ethanol or Biodiesel) Fuels : A Review Nu. *Energy & Fuels*, *21*(4), 2433-2445.

[80] Suppes, G. J., Goff, M., Burkhart, M. L., & Bockwinkel, K. (2001). Multifunctional Diesel Fuel Additives from Triglycerides. *Energy & Fuels*, *15*(1), 151-157.

[81] Conte, E., Martinho, A., Matos, H. A., & Gani, R. (2008). Combined Group-Contribution and Atom Connectivity Index-Based Methods for Estimation of Surface Tension and Viscosity. *Industrial & Engineering Chemistry Research*, 47(20), 7940-7954.

[82] Hall, L. H., & Kier, L. B. (2001). Issues in representation of molecular structure the development of molecular connectivity. *Journal of molecular graphics & modelling*, *20*(1), 4-18.

[83] Lapuerta, M., Rodríguez-Fernández, J., & De Mora, E. F. (2009). Correlation for the estimation of the cetane number of biodiesel fuels and implications on the iodine number. *Social Sciences*, *37*(11), 4337-4344.

[84] Socrates, G. (2001). *Infrared and Raman characteristic group frequencies: tables and charts*. (G. Socrates. Ed.) *Journal of Raman Spectroscopy* (Vol. 35. p. 347). Wiley.

[85] Krevelen, D. W. V., & Nijenhuis, K. T. (2009). *Properties of Polymers. Their correlation with chemical structure; their numerical estimation and prediction from additive group contributions. (Null, Eds.)Endeavour* (Vol. 16, pp. 656-658).

[86] Ayala, A. E., Simoni, L. D., Lin, Y., & Brennecke, J. F. (2006). Process design using ionic liquids : Physical property modeling. *Computer Aided Chemical Engineering*, 21, 463-468.

[87] Turner, E. A., Pye, C. C., & Singer, R. D. (2003). Use of ab Initio Calculations toward the Rational Design of Room Temperature Ionic Liquids. *Journal of Physical Chemistry*, (Table 1), 2277–2288.

[88] Holbrey, J. D., & Seddon, K. R. (1999). Ionic Liquids. *Clean Products and Processes*, *1*, 223-236.

[89] Ionic Liquids Database- (ILThermo). NIST Standard Reference database # 147. Retrieved from http://ilthermo.boulder.nist.gov/ILThermo/mainmenu.uix

[90] Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E., & Hutchison, G. R. (2012). Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of cheminformatics*, *4*(1), 17.

[91] Halgren, T. A. (1996). Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6), 490–519.

[92] Gaussian 09 (2009), Revision A.1, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian, Inc., Wallingford CT.

[93] Talaty, E. R., Raja, S., Storhaug, V. J., Do, A., & Carper, W. R. (2004). Raman and Infrared Spectra and ab Initio Calculations of C 2-4 MIM Imidazolium Hexafluorophosphate Ionic Liquids. *Journal of Physical Chemistry*, *108,* 13177–13184.

[94] Scott, A. P., & Radom, L. (1996). Harmonic Vibrational Frequencies: An Evaluation of Hartree−Fock, Møller−Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *Journal of Physical Chemistry*, *100*(41), 16502–16513.

[95] http://cccbdb.nist.gov/vibscalejust.asp

[96] Wasserscheid, P., & Welton, T. (Eds.). (2007). *Ionic Liquids in Synthesis* (2nd ed., p. 776). WILEY-VCH.

[97] Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*. (R. Mannhold, H. Kubinyi, & G. Folkers, Eds.) (2nd ed., p. 1257). Weinheim: WILEY-VCH.

[98] ADMET Predictor (2011). *User Manual*. Simulations Plus, Lancaster, CA.

[99] Bro, R., & Kiers, H. A. L. (2003). A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics*, *17*(5), 274–286.

[100] Thygesen, J., & Van Den Berg, F. (2011). Calibration transfer for excitation-emission fluorescence measurements. *Analytica Chimica Acta*, *705*(1-2), 81-7.

[101] Bieroza, M., Baker, A., & Bridgeman, J. (2011). Classification and calibration of organic matter fluorescence data with multiway analysis methods and artificial neural networks: an operational tool for improved drinking water treatment. *Environmetrics*, *22*(3), 256–270.

[102] Albrecht, T., Papadokonstantakis, S., Sugiyama, H., & Hungerbühler, K. (2010). Demonstrating multi-objective screening of chemical batch process alternatives during early design phases. *Chemical Engineering Research and Design*, *88*(5-6), 529–550.

[103] Srinivasan, R., & Nhan, N. T. (2008). A statistical approach for evaluating inherent benign-ness of chemical process routes in early design stages. *Process Safety and Environmental Protection*, *86*(3), 163–174.

[104] Banimostafa, A., Papadokonstantakis, S., & Hungerbühler, K. (2012). Evaluation of EHS hazard and sustainability metrics during early process design stages using principal component analysis. *Process Safety and Environmental Protection*, *90*(1), 8–26.

APPENDIX A

MIXTURE MODELS

The Scheffe canonical models and Cox polynomial models are the two most commonly used models to analyze mixture data with multiple regressions.

A.1 Scheffe Mixture Model

Scheffe[1] introduced canonical models of various orders by eliminating some terms from the complete polynomial model. More details can be found in Cornell[2]. The first- and second- order point estimate forms of the Scheffe models are represented by Eqs (A.2) and (A.3) respectively.

$$y = \sum_{i=1}^{n} \beta_i x_i \qquad \textbf{(A.2)}$$

$$y = \sum_{i=1}^{n} \beta_i x_i + \sum_{i<j}^{n} \sum_{j>i}^{n} \beta_{ij} x_i x_j \qquad \textbf{(A.3)}$$

By removing the constant term, the primary collinearity introduced by Eq. (3.1) is eliminated; however, it makes it impossible to center these models, which leads to ill-conditioning of the $\mathbf{X}^T\mathbf{X}$ matrix and poor estimates of the coefficients[3]. Moreover, the Scheffe model is difficult to interpret when the objective of the data analysis is determination of the component effects on

1 Scheffe, H. (1963). Simplex-centroid design for experiments with mixtures. *J. R. Stat. Soc. B.*, 25 (2), 235-263.

2 Cornell, J. (1990), Experiments with mixtures, 2nd Ed., Wiley, New York.

3 Kettaneh-Wold, N. (1992). Analysis of mixture data with partial least squares. *Chemometrics and Intelligent Laboratory Systems,* 14(1-3), 57-69.

the response. Moreover, secondary collinearity from additional constraints such as upper and lower limits on components, results in poor estimates of the regression coefficients.

A.1 Cox Mixture Model

Cox[3], recognizing the difficulties with the Scheffe mixture model, derived new mixture models with respect to a specified reference point in the experimental region. They are regular polynomials with constraints involving the reference mixture. The first and second order polynomial forms are represented by Eqs (A.4) and (A.5).

$$y = \beta_o + \sum_{i=1}^{n} \beta_i x_i \qquad \textbf{(A.4)}$$

$$y = \beta_o + \sum_{i=1}^{n} \beta_i x_i + \sum_{i \le j}^{n} \sum_{j}^{n} \beta_{ij} x_i x_j \qquad \textbf{(A.5)}$$

In terms of change in constituent i, Δ_i, Eqs (A.4) and (A.5) can be rewritten as:

$$y = y(s) + \sum_{i=1}^{n} \left(\frac{\beta_i}{1 - s_i} \right) \Delta_i \qquad \textbf{(A.6)}$$

$$y = y(s) + \sum_{i=1}^{n} \left(\frac{\beta_i}{1 - s_i} \right) \Delta_i + \sum_{i=1}^{n} \left(\frac{\beta_{ii}}{\left(1 - s_i\right)^2} \right) \Delta_i^2 \qquad \textbf{(A.7)}$$

where, *y(s)* is the expected response at the standard reference mixture.

The Cox coefficients represent the change in the response as one moves away from the standard reference mixture, and hence is meaningful in most applications. However, for multiple regressions, the Cox model encounters estimation difficulties as additional constraints are involved.

In most practical cases, (except when **X** is generated according to an experimental design), however, the **X**-variables are not statistically independent. This situation is referred to as **X** being rank deficient. Although the Scheffe canonical models and the Cox polynomial models (a reparameterized and constrained version of the Scheffe model) eliminated the *true* collinearity, and enabled the use of multiple regressions for the estimation, the problem of *near* collinearities with mixture data remains. Design of experiments (DOE) with response surface methods are usually used to determine the optimum combination of chemical constituents that give a desired response using a minimum number of experimental runs. While such a design approach is adequate for most experimental designs, it suffers from combinatorial explosion and visualization difficulties when dealing with multi-component mixtures[4]. Solvason et al.[5] presented a solution to these problems by integrating the property clustering framework with existing mixture design techniques.

4 Eden, M. R. (2003). Property-Based Process and Product Synthesis and Design. CAPEC, Department of Chemical Engineering, Technical University of Denmark. *Ph.D Thesis*.

5 Solvason, C. C., Chemmangattuvalappil, N. G., Eljack, F. T., & Eden, M. R. (2009). Efficient Visual Mixture Design of Experiments using Property Clustering Techniques. *Industrial & Engineering Chemistry Research*, 48(4), 2245-2256.

APPENDIX B

SPECTRAL INTERPRETATION

The vibrational spectrum of a molecule is considered to be a unique physical property and is a characteristic of the molecule. Any spectrum originates from radiation energy transferred to mechanical energy associated with the motion of atoms held together by chemical bonds in a molecule. The first principles approach, which is based on the fact that structural features of the molecule, whether they are the backbone of the molecule or the functional groups attached to the molecule, produce characteristic and reproducible absorptions in the spectrum[6]. This information can indicate whether there is a backbone to the structure and, if so, whether the backbone consists of linear or branched chains. Next it is possible to determine if there is unsaturation and/or aromatic rings in the structure. Finally, it is possible to deduce whether specific functional groups are present. If detected, one is also able to determine the local orientation of the group and its local environment and/or location in the structure[7].

6 Salzer, R. (2008). Practical Guide to Interpretive Near-Infrared Spectroscopy. By Jerry Workman, Jr. and Lois Weyer. *Angewandte Chemie International Edition*,47(25), 4628-4629.

7 Pasquini, C. (2003). Review Near Infrared Spectroscopy : Fundamentals , Practical Aspects and Analytical. *Applications*. Spectroscopy, 14(2), 198-219.

B.1 Infrared Spectroscopy

The fundamental absorption frequencies (also known as group frequencies) are the key to unlocking the structure–spectral relationships of the associated molecular vibrations. An infrared spectrum is formed as a consequence of the absorption of electromagnetic radiation at frequencies that correlate to the vibration of specific sets of chemical bonds from within a molecule. Figure B.6.5[8] shows the IR region of the electromagnetic spectrum. The distribution of energy possessed by a molecule at any given moment can be defined as the sum of the contributing energy terms:

$$E_{total} = E_{electronic} + E_{vibrational} + E_{rotational} + E_{trnaslational} \quad \textbf{(B.1)}$$

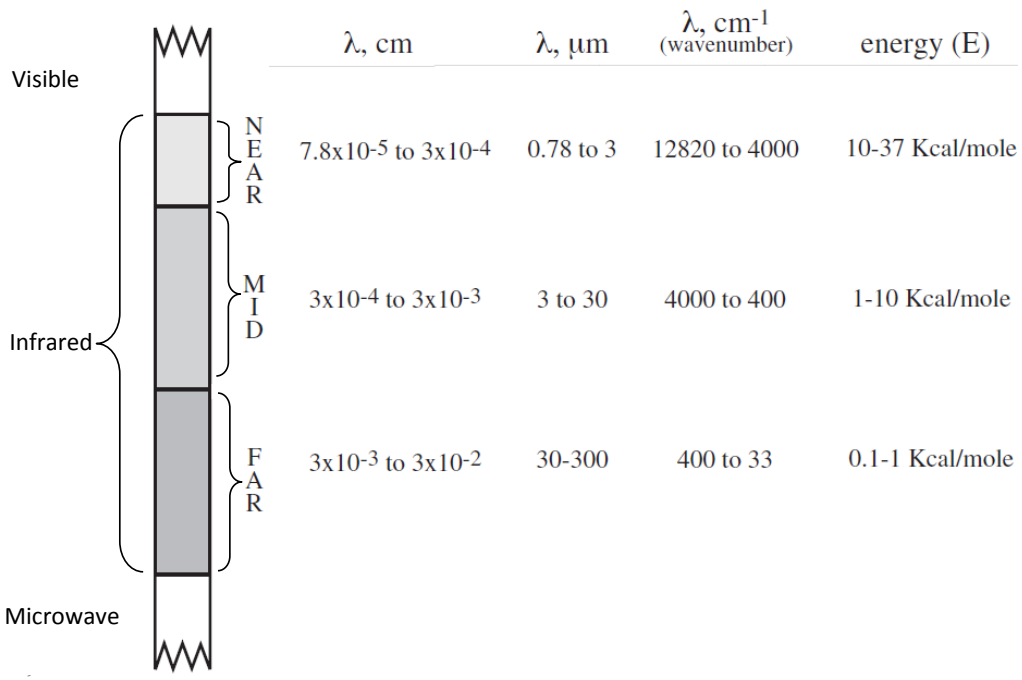


Figure B.6.5: IR regions of the electromagnetic spectrum.

8 Raffael, K. D. (2002). Infrared Spectroscopy: The Theory. *Journal of Molecular Spectroscopy*, 214(1), 21-27

B.2 Molecular Vibrational Spectroscopy

There are two types of molecular vibrations, stretching and bending. A molecule consisting of n-atoms has $3n$ degrees of freedom, corresponding to the Cartesian coordinates of each atom in the molecule. In a nonlinear molecule, 3 of these degrees are rotational and 3 are translational and the remaining corresponds to fundamental vibrations; in a linear molecule, 2 degrees are rotational and 3 are translational. This is because in a linear molecule, all of the atoms lie on a single straight line and hence rotation about the bond axis is not possible. The net number of vibrational degrees of freedom for a given molecule can be determined from Eq. (B.2):

$$\begin{aligned} Number\ of\ normal\ mode &= 3n-6 \text{ (nonlinear)} \\ &= 3n-5 \text{ (linear)} \end{aligned} \qquad \textbf{(B.2)}$$

For example, water, which is nonlinear, has three fundamental vibrations as shown in Figure B.6.6[8].

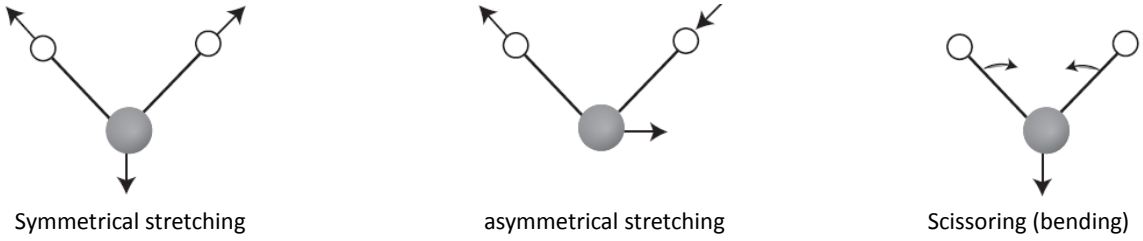


Figure B.6.6: Stretching and bending vibrational modes for H_2O.

If we calculate the number of modes for a simple hydrocarbon, such as methane (nonlinear, tetrahedral structure), a value of nine is obtained. This would imply that nine sets of absorption frequencies would be observed in the spectrum of methane gas. In practice, infrared spectra do not normally display separate absorption signals for each of the $3n$-6 fundamental vibrational modes of a molecule. The number of observed absorptions may be

increased by additive and subtractive interactions leading to combination tones and overtones of the fundamental vibrations, in much the same way that sound vibrations from a musical instrument interact. Furthermore, the number of observed absorptions may be decreased by molecular symmetry, spectrometer limitations, and spectroscopic selection rules.

One selection rule that influences the intensity of infrared absorptions is that a change in dipole moment of the molecule should occur for a vibration to absorb infrared energy. Absorption bands associated with C=O bond stretching are usually very strong because a large change in the dipole takes place in that mode. The reason for the smaller than expected number is that several of the vibrations are redundant or degenerate, that is, the same amount of energy is required for these vibrations. Some general trends are as follows:

- Stretching frequencies are higher than corresponding bending frequencies. (It is easier to bend a bond than to stretch or compress it.)
- Bonds to hydrogen have higher stretching frequencies than those to heavier atoms.
- Triple bonds have higher stretching frequencies than corresponding double bonds, which in turn have higher frequencies than single bonds. (Except for bonds to hydrogen).

The stretching and bending vibrations for the important organic group, $-CH_2$, are illustrated in Figure B.6.7[9]. (The $3n-6$ rule does not apply since the $-CH_2$ group represents only a portion of a molecule.) Note that bending vibrations occur at lower frequencies than the corresponding stretching vibrations.

9 Socrates, G. (2001). *Infrared and Raman characteristic group frequencies: tables and charts*. (G. Socrates. Ed.) Journal of Raman Spectroscopy (Vol. 35. p. 347). Wiley.

The fundamental requirement for infrared activity, leading to absorption of infrared radiation, is that there must be a net change in dipole moment during the vibration for the molecule or the functional group under study. Another important form of vibrational spectroscopy is Raman spectroscopy, which is complementary to infrared spectroscopy. The selection rules for Raman spectroscopy are different to those for infrared spectroscopy, and in this case a net change in bond polarizability must be observed for a transition to be Raman active.

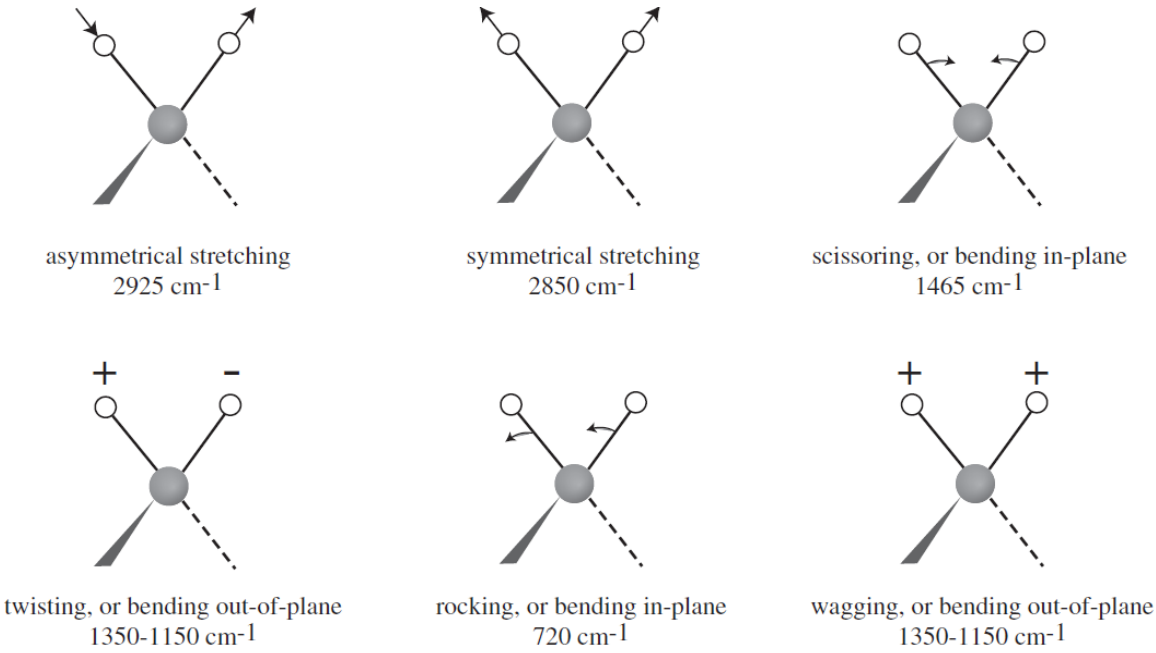


Figure B.6.7: Stretching and bending vibrational modes for a CH_2 goup.

Both the stretching and bending vibrations of a molecule as illustrated in the above figures can be predicted mathematically, at least to a useful approximation using the following mathematical description of stretching vibrations.

B.2.1 Stretching Vibration

The stretching frequency of a bond can be approximated by Hooke's Law. In this approximation, the fundamental frequency of vibration of two atoms and the connecting covalent bond are treated as a simple harmonic oscillator composed of 2 masses (atoms) joined by a spring and can be expressed by:

$$\nu = \frac{1}{2\pi}\sqrt{\frac{\kappa}{\mu}} \qquad \textbf{(B.3)}$$

Where ν is the vibrational frequency

κ is the force constant of the spring, and

μ is the reduced mass.

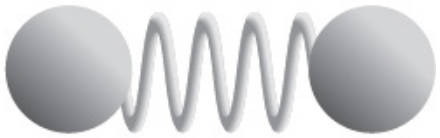


Figure B.6.8: Ball and spring model for atoms and bonds respectively.

Using vibrations and wave equations of quantum mechanics, the potential energy can be written as:

$$E = (n + \frac{1}{2})h\nu \qquad \textbf{(B.4)}$$

where, h is the Plank's constant, 6.6 x 10^{-34} J/s,

n is the vibrational quantum number (0, 1, 2, 3, . . .)

At the ground state ($\nu = 0$) $E_0 = 1/2\,h\nu$. Following the selection rule, when a molecule absorbs energy, there is a promotion to the first excited state $E_1 = 3/2\,h\nu$. These correspond to bands called overtones in an IR spectrum.

They are of lower intensity than the fundamental vibration bands. The difference in energy levels between the vibrational quantum states exactly matches the radiation energy expressed as:

$$\Delta E = \left(\frac{3}{2}h\nu - \frac{1}{2}h\nu\right) = h\nu = \frac{h}{2\pi}\sqrt{\frac{\kappa}{\mu}} \qquad \textbf{(B.5)}$$

Since, $\nu = c/\lambda = c\bar{\nu}$, and Eq. (B.3) for a diatomic molecule becomes

$$\bar{\nu} = \frac{1}{2\pi c}\sqrt{\frac{\kappa\left(m_1 + m_2\right)}{m_1 m_1}} \qquad \text{where, } \mu = \frac{m_1 \cdot m_2}{m_1 + m_2} \qquad \textbf{(B.6)}$$

where, c is the speed of light, 3 x 10^8m/s.

$\bar{\nu}$ is the wavenumber, inverse of the wavelength, λ.

For example, using Hooke's law approximation, C-H bond stretching vibrations can be estimated as:

$$m_C := \frac{12gm}{6.023\cdot 10^{23}} \qquad m_H := \frac{1gm}{6.023\cdot 10^{23}}$$

$$\nu := \frac{1}{2\pi\cdot c}\sqrt{\frac{\kappa\cdot\left(m_C + m_H\right)}{m_C\cdot m_H}}$$

$$\boxed{\nu = 3032\,cm^{-1}}$$

The actual range for C–H absorptions is 2850–3000 cm^{-1}. The region of an IR spectrum where bond stretching vibrations is seen depends primarily on whether the bonds are single, double, or triple or bonds to hydrogen. The following table shows where absorption by single, double, and triple bonds are observed in an IR spectrum.

Table B.1: IR spectrum absorption for different bond types.

Bond type	Force constant dyne/cm	Bond	IR absorption range cm^{-1}
Single	5×10^5	C-C, C-O, C-N	800-1300
Double	10×10^5	C=C, C=O, C=N, N=O	1500-1900
Triple	15×10^5	C≡C, C≡N	2000-2300
Hydrogen		C-H, N-H, O-H	2700-3800

The general regions of the infrared spectrum in which various kinds of vibrational bands are observed are outlined in the Figure B.6.9.

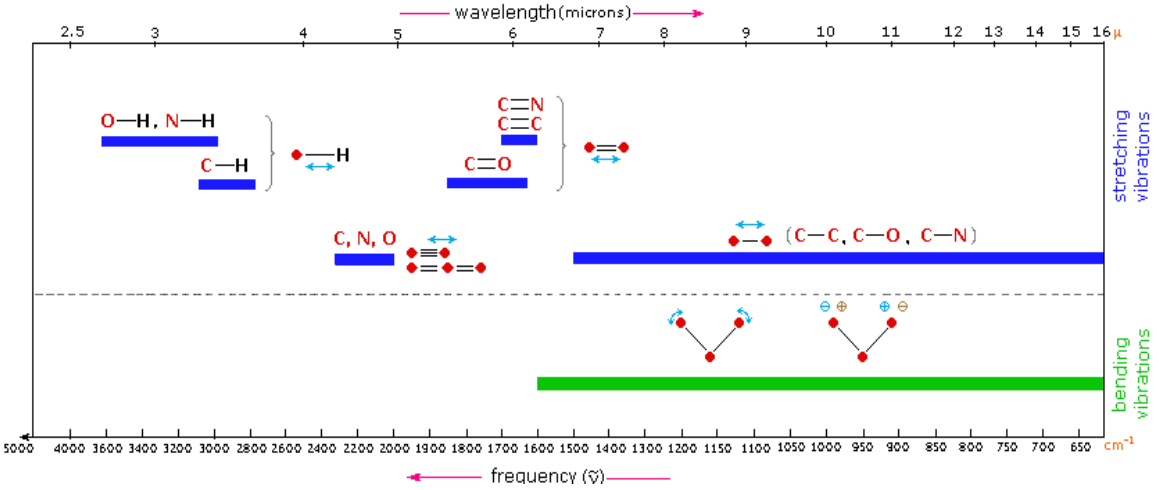


Figure B.6.9: Vibrational bands in infrared spectrum.

B.3 Near Infrared Spectroscopy

Absorption bands in the near infrared region (NIR) (12820 - 4000 cm^{-1}) are weak because they arise from vibrational overtones and combination bands. Combination bands occur when two molecular vibrations are excited simultaneously. The intensity of overtone bands reduces by one order of overtone for each successive overtone when a molecule is excited from the ground vibrational state to a higher vibrational state. When the vibrational

quantum number υ is greater than or equal to 2 then an overtone absorption results. The first overtone results from $\nu = 0$ to $\nu = 2$ $(5/2\,h\nu - 1/2\,h\nu = 2h\nu)$. The second overtone occurs when $\nu = 0$ transitions to $\nu = 3$ $\left(7/2\,h\nu - 1/2\,h\nu = 3h\nu\right)$. Transitions arising from the near-IR absorption are weak, hence they are referred to as forbidden transitions but these transitions are relevant when nondestructive measurements are required such as a solid sample.

B.4 Characterizing IR Spectroscopy

Since relative intensities are primarily functions of the atom specific dipole changes caused by the vibration of the corresponding bonds, it follows that their size and shape are indicators of molecular architecture. The IR absorbance frequencies and magnitudes of the functional groups spectrums are listed in following tables and were compiled from Socrates[9].

Table B.2: IR absorbance frequencies and magnitudes of functional groups.

Methine Groups, -CH- -

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
Bending *(δ)*	1360	1320	1340	*w*	90
Stretching *(υ)*	2890	2880	2885	*w*	90

Methylene Groups, -CH$_2$-

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
Scissoring Bend (δ_s)	1480	1440	1460	*m*	50
Symmetrical Stretching (υ_s)	2870	2840	2855	*m*	50
Asymmetrical Stretching (υ_a)	2940	2915	2928	*m-s*	30

Methyl Groups, -CH_3

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
Sym Bend (δ_s)	1390	1370	1380	*m-s*	30
Asym. Bend. (δ_a)	1465	1440	1453	*m*	50
Symmetrical Stretching (υ_s)	2885	2865	2875	*m*	50
Asymmetrical Stretching (υ_a)	2975	2950	2963	*m-s*	30

Tetramethyl Groups, -$C(CH_3)_3$

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-C Skeletal Bend (δ_s)	930	925	928	*m*	50
C-C Skeletal Bend (δ_s)	1010	990	1000	*m-w*	70
C-C Skeletal Bend (δ_s)	1225	1165	1195	*m*	50
C-C Skeletal Bend (δ_s)	1255	1245	1250	*m*	50
C-CH_3 Sym. Bend. (δ_s)	1395	1350	1373	*m-s*	30
C-CH_3 Sym. Bend. (δ_s)	1420	1375	1398	*m*	50
C-CH_3 Asym. Bend. (δ_a)	1475	1435	1455	*m*	50
C-H Sym. Stretching (υ_s)	2885	2865	2875	*m*	50
C-H Asym. Stretching (υ_a)	2975	2950	2963	*m-s*	30

Aliphatic Methoxy Groups, -O-CH_3 (Special Methyl)

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-O Def. Bend. (δ_d)	580	340	460	*m-w*	70
CH_3/CO Rocking Bend (δ_d)	1190	1100	1145	*m-w*	70
CH_3 Rock Bend (δ_d)	1235	1155	1195	*m-w*	70
CH_3 Sym Bend (δ_s)	1460	1420	1440	*M*	50
CH_3 Asym. Bend. (δ_a)	1475	1435	1455	*m*	50
CH_3 Asym. Bend. (δ_a)	1485	1445	1465	*m*	50
C-H_3 Sym. Str. (υ_s)	2880	2815	2848	*m*	50

C-H_3 Asym. Str. (υ_a)	2985	2920	2953	*m*	50
C-H Asym. Str. (υ_a)	3030	2950	2990	*m*	50

Vinyl Group, -CH=CH_2

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C=C Tors. Bend (δ_T)	485	410	448	*m-s*	30
C=C Eth. Twist. Bend. (δ_t)	600	380	490	*m-s*	30
C=C Eth. Twist. Bend. (δ_t)	720	410	565	*w*	90
C-H_2 OoP Rock. Bend. (δ_r)	980	810	895	*s*	10
C-H OoP Bending. (δ_r)	1010	940	975	*s*	10
C-H IP Def. Bend. (δ_d)	1180	1010	1095	*m-w*	70
C-H_2 Def. Bend. (δ_d)	1330	1240	1285	*m*	50
C-H_2 Sci. Bend. (δ_s)	1440	1360	1400	*m*	50
C=C Stretching (υ)	1645	1640	1643	*m-w*	70
C-H_2 1st Overtone Bend (2δ)	1840	1820	1830	*v*	90
C-H 1st Overtone Bend (2δ)	1990	1970	1980	*v*	90
C-H_2 Sym. Stretch (υ_s)	3070	2930	3000	*M*	50
C-H Stretch (υ)	3110	2980	3045	*M*	50
C-H_2 Asym. Stretch (υ_a)	3150	3000	3075	*M*	50

Vinylidene Group, CH_2=C- -

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C=C Skeletal Stretch (υ)	470	435	453	*m-w*	70
C=C Skeletal Stretch (υ)	560	530	545	*s*	10
C=C Eth. Twist. Bend. (δ_t)	715	680	698	*w*	90
C-H_2 OoP Rock. Bend. (δ_r)	895	885	890	*s*	10
C-H_2 IP Def. Bend. (δ_d)	1320	1290	1305	*w*	90
C-H_2 Sci. Def Bend. (δ_s)	1420	1405	1413	*w*	90

C=C Stretching (υ)	1675	1625	1650	*m-w*	70
C-H_2 1st Overtone Bend (2δ)	1800	1750	1775	*w*	90
C-H_2 Sym. Stretch (υ_s)	2985	2970	2978	*m-w*	70
C-H_2 Asym. Stretch (υ_a)	3095	3075	3085	*m-w*	70

cis-Vinylene Group, -CH=CH-

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-H Tors. Bend (δ_T)	490	320	405	*m-s*	30
C=C Skeletal Bend (δ_T)	500	460	480	*s*	10
-C=CH Def. Bend. (δ_d)	590	440	515	*m-s*	30
C=C Eth. Twist. Bend. (δ_t)	630	570	600	*s*	10
C-H Wag. Bend. (δ_w)	790	650	720	*m-s*	30
C-H Wag. Bend. (δ_w)	1000	850	925	*m-w*	70
C-H Def. Bend. (δ_d)	1295	1185	1240	*w*	90
C-H Def. Bend. (δ_d)	1425	1355	1390	*w*	90
C=C Stretching (υ)	1665	1630	1648	*m*	50
C-H Stretch (υ)	3040	2980	3010	*m*	50
C-H Stretch (υ)	3090	3010	3050	*m*	50

trans-Vinylene Group, -CH=CH-

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-H Tors. Bend (δ_T)	490	320	405	*m-s*	30
C=C Skeletal Bend (δ_T)	500	480	490	*s*	10
-C=CH Def. Bend. (δ_d)	590	440	515	*m-s*	30
C=C Eth. Twist. Bend. (δ_t)	580	515	548	*m-s*	30
C-H Wag. Bend. (δ_w)	850	750	800	*m-w*	70
C-H Wag. Bend. (δ_w)	1000	910	955	*v*	90
C-H Def. Bend. (δ_d)	1305	1260	1282.5	*v*	90
C-H Def. Bend. (δ_d)	1340	1355	1347.5	*v*	90

C=C Stretching (υ)	1680	1665	1673	*m-w*	70
C-H Stretch (υ)	3050	3000	3025	*m*	50
C-H Stretch (υ)	3065	3015	3040	*m*	50

Hydroxyl Group, -OH (with intermolecular H-bonding)

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
Bending (δ)	710	570	640	*m*	50
Stretching (υ)	3550	3230	3390	*m-s*	30

Primary Alcohol Group, -CH_2OH (with intermolecular H-bonding)

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-O Def. Bend (δ_d)	555	395	475	*m-w*	70
C-O IP. Def. Bend (δ_d)	500	440	470	*w*	90
O-H OoP. Def. Bending (δ_d)	710	570	640	*m-w*	70
C-CO Stretch (υ)	900	800	850	*m*	50
C-H_2 Twist Bend (δ_t)	960	800	880	*m-w*	70
C-C-O Stretch (υ)	1090	1000	1045	*S*	10
C-H_2 Twist. Bending (δ_t)	1300	1280	1290	*m-w*	70
C-H_2 Wag Bend (δ_w)	1390	1280	1335	*m-w*	70
O-H Def. Bend (δ_d)	1440	1260	1350	*m-s*	30
C-H_2 Def Bend (δ_d)	1480	1410	1445	*m-w*	70
C-H_2 Sym. Stretch (υ_s)	2935	2840	2888	*m-w*	70
C-H_2 Asym. Stretch (υ_a)	2990	2900	2945	*m-w*	70
O-H Stretching (υ)	3550	3230	3390	*m-s*	30

Secondary Alcohol Group, - -CHOH (with intermolecular H-bonding)

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-O OoP. Def. Bend (δ_d)	390	330	360	*m-w*	70
C-O IP. Def. Bend (δ_d)	500	440	470	*w*	90
O-H OoP. Def. Bending (δ_d)	660	600	630	*m-w*	70
C-CO Stretch (υ)	900	800	850	*m*	50
C-O Stretch (υ)	1150	1075	1113	*m-w*	70
C-H Def. Bending (δ_d)	1350	1290	1320	*s*	10
C-H Wag Bend (δ_w)	1400	1330	1365	*s*	10
O-H + C-H_2 Coup. Bend. (δ_c)	1430	1370	1400	*m-w*	70
O-H Def. Bend (δ_d)	1440	1260	1350	*m-w*	70
C-H Stretching (υ)	2890	2880	2885	*m-s*	30
O-H Stretching (υ)	3550	3230	3390	*m-w*	70

Aliphatic Ether Group, -O-

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-O-C def vib (δ_d)	440	420	430	*w*	90
Sym C-O-C str (υ_s)	1140	820	980	*w*	90
Asym C-O-C Str (υ_a)	1150	1060	1105	*s*	10
Rocking vib	1200	1185	1193	*m-w*	70
Wagging vib	1400	1360	1380	*m*	50
Asym and Sym -CH_3 def. vib	1470	1435	1453	*m*	50
CH_2 def vib	1475	1445	1460	*m*	50
Sym CH_2 str	2880	2835	2858	*m*	50
Sym. -CH_3 Str	2900	2840	2870	*m*	50
Asym CH_2 str	2955	2920	2938	*m*	50
Asym. -CH_3 Str	2995	2955	2975	*m*	50

Alkyl Peroxide Group, -O-O-

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
O-O Stretch (υ)	900	800	850	*w*	90
C-O Stretch (υ)	1150	1030	1090	*m-s*	30

Saturated Aliphatic Ester Group, -CO-O-

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-O-C Sym. Stretch (υ_s)	1160	1050	1105	*s*	10
C-O-C Asym. Stretch (υ_a)	1275	1185	1230	*s*	10
C=O Stretch (υ)	1750	1725	1738	*s*	10
C=O 1st Overtone ($2\upsilon_s$)	3460	3440	3450	*w*	90

Saturated Aliphatic Methyl Ester Group, -CO-O-CH_3

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
Unlisted	450	430	440	*m-s*	30
CO-O Rocking Bend (δ_r)	530	340	435	*w*	90
C-C-O Sym. Stretch (υ_s)	1160	1050	1105	*s*	10
C-O Stretch (υ)	1175	1155	1165	*s*	10
C-C-O Asym. Stretch (υ_a)	1275	1185	1230	*s*	10
O-CH_3 Stretch (υ)	1315	1195	1255	*s*	10
Unlisted	1370	1350	1360	*w*	90
CH_3 Sym. Def. Bend (δ_d)	1460	1420	1440	*m-s*	30
CH_3 Asym. Def. Bend (δ_d)	1465	1420	1443	*m-s*	30
CH_3 Asym. Def. Bend (δ_d)	1485	1435	1460	*m*	50
C=O Stretch (υ)	1750	1725	1738	*s*	10
CH_3 Sym. Stretch (υ)	3000	2860	2930	*m*	50

CH_3 Asym. Stretch (υ)	3030	2950	2990	*m-w*	70
CH_3 Asym. Stretch (υ)	3050	2980	3015	*m-w*	70
C=O 1st Overtone ($2\upsilon_s$)	3460	3440	3450	*w*	90

Saturated Aliphatic Ethyl Ester Group, -CO-O-CH_2CH_3

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-O-C Def Bend (δ_d)	370	250	310	*m-w*	70
C-O-C Def Bend (δ_d)	395	305	350	*m-w*	70
CO-O Rocking Bend (δ_r)	485	365	425	*m-w*	70
CO OoP Rocking Bend (δ_r)	700	550	625	*w*	90
CH_2 Rocking Bend (δ_r)	825	775	800	*w*	90
C-C str (υ)	940	850	895	*w*	90
CH_3 Rock. Bend (δ_r)	1150	1080	1115	*w*	90
C-C-O Sym. Stretch (υ_s)	1160	1050	1105	*s*	10
CH_3 Rock. Bend (δ_r)	1195	1135	1165	*w*	90
C-C-O Asym. Stretch (υ_a)	1275	1185	1230	*s*	10
CH_2 Twist. Bend (δ_T)	1340	1325	1333	*m-w*	70
CH_2 Wag. Bend (δ_w)	1385	1335	1360	*m-w*	70
CH_3 Sym. Def. Bend (δ)	1390	1360	1375	*m-s*	30
CH_3 Asym. Def. Bend (δ)	1480	1435	1458	*m*	50
OCH_2 Def. Bend. (δ)	1490	1460	1475	*m-w*	70
C=O Stretch (υ)	1750	1725	1738	*s*	10
CH_3 Stretch (υ)	2920	2860	2890	*w*	90
CH_3 Sym. Stretch (υ_s)	2930	2890	2910	*w*	90
CH_3 Asym. Stretch (υ_a)	2995	2930	2963	*m*	50
C=O 1st Overtone ($2\upsilon_s$)	3460	3440	3450	*w*	90

Acrylate Ester Group, CH_2=CH-CO-O-

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C=C Tors. Bend (δ_T)	485	410	448	*m-s*	30
CO-O Rocking Bend (δ_r)	485	365	425	*m-w*	70
C=C Eth. Twist. Bend. (δ_t)	600	380	490	*m-s*	30
C-O-C Def Bend (δ)	675	660	668	*m*	50
CO OoP Rocking Bend (δ_r)	700	550	625	*w*	90
=CH2 Twist Bend (δ_t)	810	800	805	*m-s*	30
CH_2 Rocking Bend (δ_r)	825	775	800	*w*	90
C-C str (υ)	940	850	895	*w*	90
=CH2 Wag. Bend (δ_w)	970	960	965	*s*	10
C-H Def. Wag (δ_w)	990	980	985	*m*	50
C-H OoP Bending. (δ_r)	1010	940	975	*s*	10
C-C Skel. Bend (δ)	1070	1065	1068	*m*	50
CH_3 Rock. Bend (δ_r)	1150	1080	1115	*w*	90
C-C-O Sym. Stretch (υ_s)	1160	1050	1105	*s*	10
C-H IP Def. Bend. (δ_d)	1180	1010	1095	*m-w*	70
Unlisted	1200	1195	1198	*s*	10
C-C-O Asym. Stretch (υ_a)	1275	1185	1230	*s*	10
=CH Rock. Bend (δ_r)	1290	1270	1280	*m*	50
Unlisted	1290	1280	1285	*s*	10
=CH2 Def Bend (δ)	1420	1400	1410	*m*	50
C-H_2 Sci. Bend. (δ_s)	1440	1360	1400	*m*	50
C=C Stretch (υ)	1635	1615	1625	*m*	50
C=C Stretch (υ)	1650	1630	1640	*m-s*	30
C=O Stretch (υ)	1725	1710	1718	*s*	10
C-H_2 1st Overtone Bend (2δ)	1840	1820	1830	*w*	90
C-H 1st Overtone Bend (2δ)	1990	1970	1980	*w*	90
C-H_2 Sym. Stretch (υ_s)	3070	2930	3000	*m*	50
C-H Stretch (υ)	3110	2980	3045	*m*	50
C-H_2 Asym. Stretch (υ_a)	3150	3000	3075	*m*	50
C=O 1st Overtone ($2\upsilon_s$)	3460	3440	3450	*w*	90

Methacrylate Ester Group, $CH_2=C(CH_3)-CO-O-$

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C=C Skeletal Stretch (υ)	470	435	453	*m-w*	70
C=C Skeletal Stretch (υ)	560	530	545	*s*	10
C-O-C Def Bend (δ)	660	645	653	*m*	50
C=C Eth. Twist. Bend. (δ_t)	715	680	698	*w*	90
C-C Skel Bend (δ)	825	805	815	*m-s*	30
$C-H_2$ OoP Rock. Bend. (δ_r)	895	885	890	*s*	10
=CH2 Wag. Bend (δ_w)	950	935	942.5	*s*	10
C-C Skel. Bend (δ)	1010	990	1000	*m*	50
C-C Skel. Bend (δ)	1020	1000	1010	*m*	50
C-O-C Sym. Stretch (υ_s)	1160	1150	1155	*s*	10
C-O-C Asym. Stretch (υ_a)	1275	1185	1230	*s*	10
Unlisted	1310	1290	1300	*s*	10
$C-H_2$ IP Def. Bend. (δ_d)	1320	1290	1305	*w*	90
=CH Rock. Bend (δ_r)	1335	1315	1325	*m*	50
CH_3 Sym Bend (δ_s)	1390	1370	1380	*m-s*	30
=CH2 Def Bend (δ)	1420	1400	1410	*m*	50
CH_3 Asym. Bend. (δ_a)	1465	1440	1453	*m*	50
C=C Stretch (υ)	1650	1630	1640	*m*	50
C=O Stretch (υ)	1725	1710	1718	*s*	10
$C-H_2$ 1st Overtone Bend (2δ)	1800	1750	1775	*w*	90
CH_3 Sym. Stretching (υ_s)	2885	2865	2875	*m*	50
$C-H_2$ Sym. Stretch (υ_s)	2985	2970	2978	*m-w*	70
CH_3 Asym. Stretching (υ_a)	2975	2950	2963	*m-s*	30
$C-H_2$ Asym. Stretch (υ_a)	3095	3075	3085	*m-w*	70
C=O 1st Overtone ($2\upsilon_s$)	3460	3440	3450	*w*	90

o-Alkyl Phenol Group (With H-bonding)

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-OH IP Bending (δ)	450	375	413	*w*	90
O-H OoP. Def. Bending (δ_d)	720	600	660	*s*	10
C-O Stretch (υ)	1260	1180	1220	*s*	10
O-H IP Bending (δ)	1410	1310	1360	*s*	10
COH bending vib	1330	1310	1320	*m*	50
O-H Stretching (υ)	3250	3000	3125	*m*	50
CO Str	1255	1240	1248	*s*	10
OH def and CO str vib	1175	1160	1168	*s*	10
OH def and CO str vib	760	740	750	*m*	50
OR substituted	3595	3470	3533	*m*	50

p-Alkyl Phenol Group (With H-bonding)

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-OH IP Bending (δ)	450	375	413	*w*	90
O-H OoP. Def. Bending (δ_d)	720	600	660	*s*	10
C-O Stretch (υ)	1260	1180	1220	*s*	10
O-H IP Bending (δ)	1410	1310	1360	*s*	10
O-H Stretching (υ)	3250	3000	3125	*m*	50
CO Str	1260	1245	1253	*s*	10
OH def and CO str vib	1175	1165	1170	*s*	10
OH def and CO str vib	835	815	825	*m*	50
OR substituted	3595	3470	3533	*m*	50

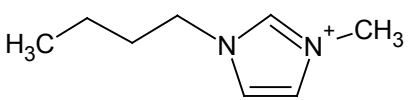
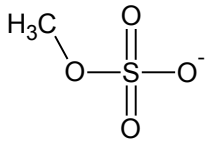
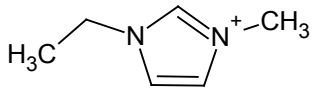
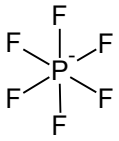
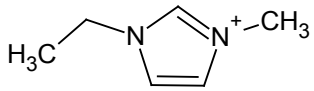
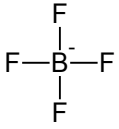
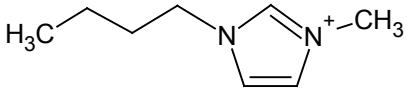
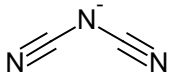
Monosubstituted Benzenes

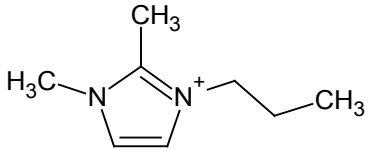
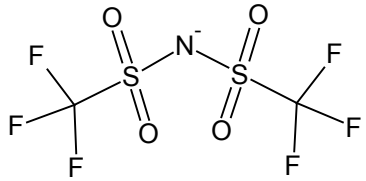
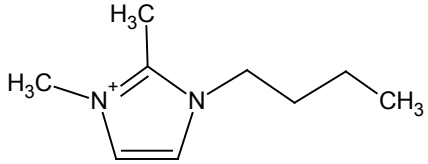
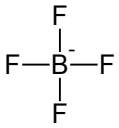
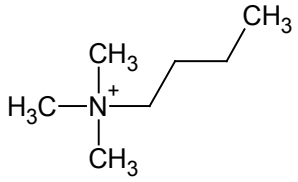
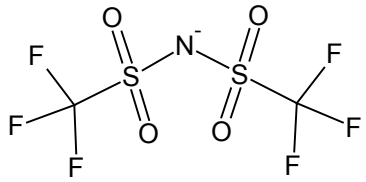
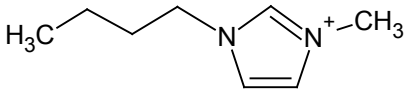
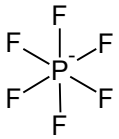
Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
Ring OoP def vib	560	415	488	*m-s*	30
Ring IP def vib	630	605	618	*m-w*	70
=C-H Ring OoP def vib	710	670	690	*s*	10
=C-H OoP def vib	820	720	770	*s*	10
=C-H OoP def vib	900	860	880	*m-w*	70
=C-H IP def vib	1010	990	1000	*w*	90
=C-H IP def vib	1040	1000	1020	*m-w*	70
=C-H IP def vib	1085	1050	1068	*m*	50
=C-H IP def vib	1175	1130	1153	*w*	90
=C-H IP def vib	1195	1165	1180	*m-w*	70
=C-H IP def vib	1250	1230	1240	*w*	90
-C=C- Str Vib	1625	1590	1608	*v*	90
=C-H Str. Vib	3105	3000	3053	*m*	50

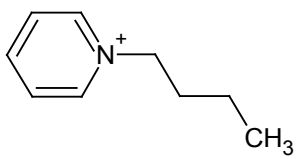
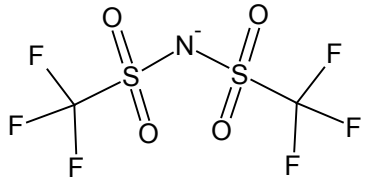
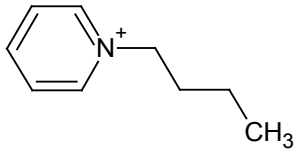
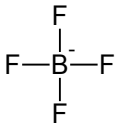
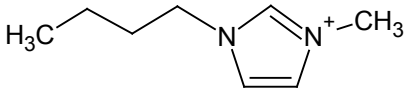
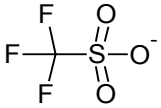
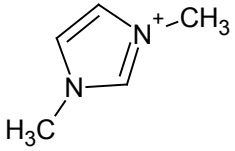
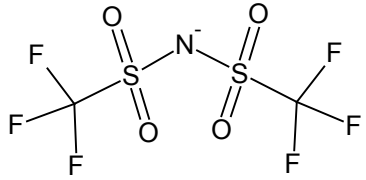
1,2,4- Trisubstituted Benzene

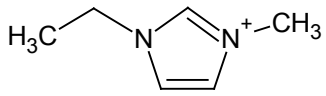
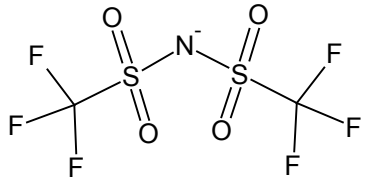
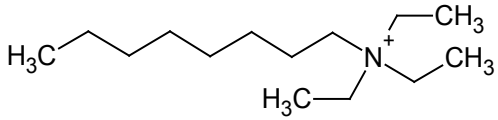
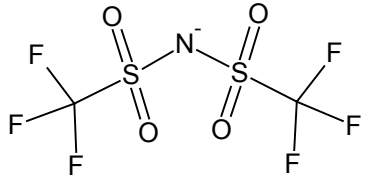
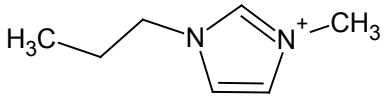
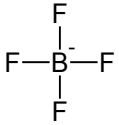
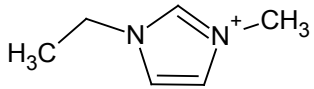
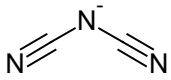
Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
Ring OoP def vib	475	425	450	*m-s*	30
=C-H OoP def vib (2H)	740	690	715	*m-w*	70
=C-H OoP def vib (2H)	780	760	770	*s*	10
=C-H OoP def vib (2H)	860	840	850	*m-s*	30
=C-H OoP def vib (1H)	940	885	913	*m-s*	30
=C-H IP def vib	1040	1020	1030	*m-w*	70
=C-H IP def vib	1160	1140	1150	*m-w*	70
=C-H IP def vib	1220	1200	1210	*w*	90
-C=C- Str Vib	1625	1590	1608	*v*	90
=C-H Str. Vib	3105	3000	3053	*m*	50

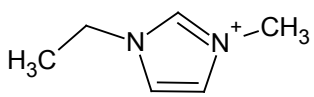
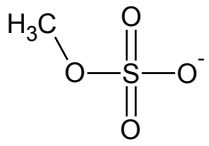
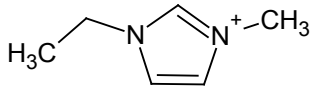
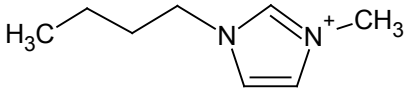
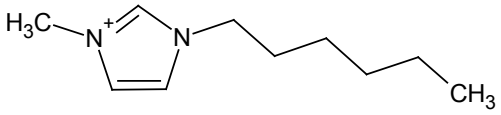
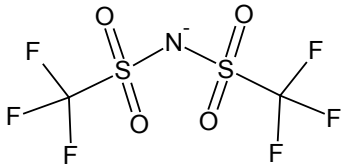
Table B.3: Abbreviations, names and structures of investigated ionic liquid training set.

ID	Abbreviation	Structure & Name		Properties		
		Cation	Anion	μ @ 293K Pa.s	ρ @ 298K Kg/m³	T_m K
1	[bmIm]mSO$_4$			0.2890	1212	269.1
		1-butyl-3-methyl imidazolium	Methyl sulfate			
2	[emIm]PF$_6$			0.0234	1422	307.1
		1-ethyl-3-methyl imidazolium	Hexafluorophosphate			
3	[emIm]BF$_4$			0.0665	1248	247.1
		1-ethyl-3-methyl imidazolium	Tetrafluoroborate			
4	[bmIm]dCN			0.0332	1058	267.1
		1-butyl-3-methyl imidazolium	Dicyanamide			

ID	Abbreviation	Structure & Name		Properties		
		Cation	Anion	μ @ 293K Pa.s	ρ @ 298K Kg/m³	T_m K
5	$[1,2m_2pIm]Tf_2N$			0.0901	1457	288.1
		1,2-dimethyl-3-propyl imidazolium	Bis(trifluromethylsulfonly)-amide			
6	$[b2,3m_2Im]BF_4$			0.4558	1094	310.1
		1-butyl-2,3-dimethyl imidazolium	Tetrafluoroborate			
7	$[bm_3NH_4]Tf_2N$			0.1407	1397	289.1
		butyl-trimethyl-ammonium	Bis(trifluromethylsulfonly)-amide			
8	$[bmIm]PF_6$			0.4500	1360	283.1
		1-butyl-3-methyl imidazolium	Hexafluorophosphate			

ID	Abbreviation	Structure & Name		Properties		
		Cation	Anion	μ @ 293K	ρ @ 298K	T_m
				Pa.s	Kg/m^3	K
9	$[bPy]Tf_2N$			0.0756	1454	299.1
		1-butylpyridinium	Bis(trifluromethylsulfonly)-amide			
10	$[bPy]BF_4$			0.2231	1203	279.8
		1-butylpyridinium	Tetrafluoroborate			
11	$[bmIm]CF_3SO_3$			0.0990	1384	262.2
		1-butyl-3-methyl imidazolium	Triflate			
12	$[1,3m_2Im]Tf_2N$			0.0475	1570	299.1
		1,3-dimethyl imidazolium	Bis(trifluromethylsulfonly)-amide			

ID	Abbreviation	Structure & Name		Properties		
		Cation	Anion	μ @ 293K	ρ @ 298K	T_m
				Pa.s	Kg/m^3	K
13	$[m_2Im]Tf_2N$			0.0260	1519	240.1
		1-ethyl-3-methyl imidazolium	Bis(trifluromethylsulfonly)-amide			
14	$[e_3oNH_4]Tf_2N$			0.1810	1270	287.1
		N,N,N-triethyl-1-octanaminium	Bis(trifluromethylsulfonly)-amide			
15	$[pmIm]PF_6$			0.1030	1240	256.1
		1-propyl-3-methyl imidazolium	Tetrafluoroborate			
16	[emIm]dCN			0.0169	1106	255
		1-ethyl-3-methyl imidazolium	Dicyanamide			

ID	Abbreviation	Structure & Name		Properties		
		Cation	Anion	μ @ 293K	ρ @ 298K	T_m
				Pa.s	Kg/m^3	K
17	[emIm]mSO_4			0.0785	1234	236.3
		1-ethyl-3-methyl imidazolium	Methylsulfate			
18	[emIm]Cl		Cl^-	1.583	1186	358.1
		1-ethyl-3-methyl imidazolium	Chloride			
19	[bmIm]Cl		Cl^-	40.89	1080	340.1
		1-butyl-3-methyl imidazolium	Chloride			
20	[hmIm]Tf_2N			0.0780	1370	266
		1-hexyl-3-methylimidazolium	Bis(trifluromethylsulfonly)-amide			

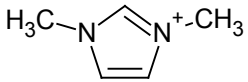
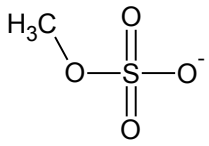
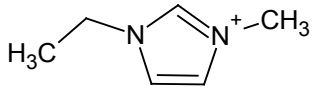
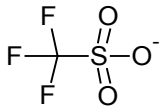
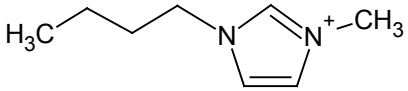
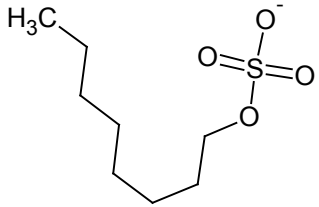
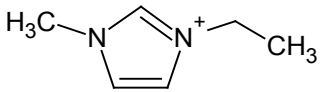
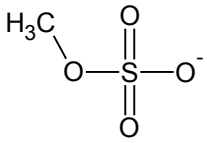
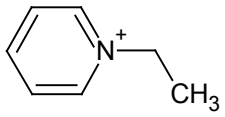
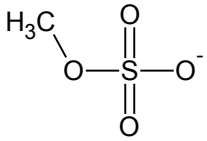
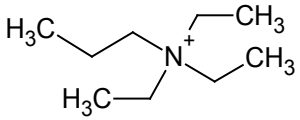
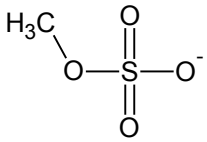
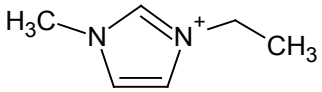
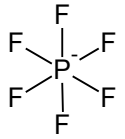
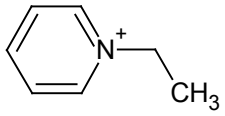
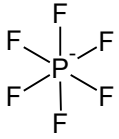
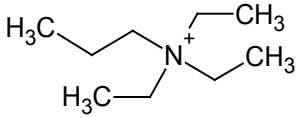
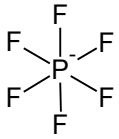
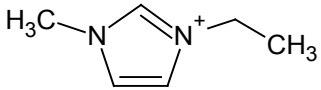
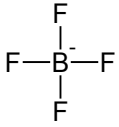
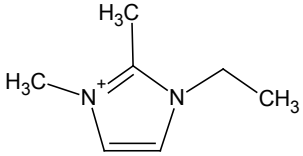
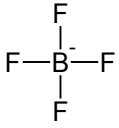
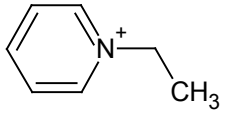
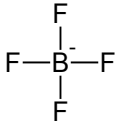
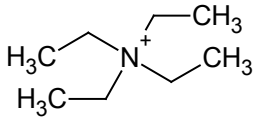
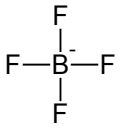
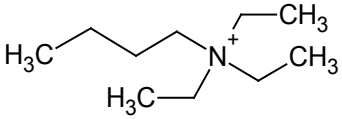
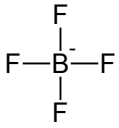
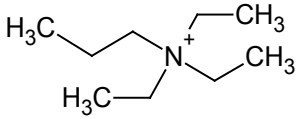
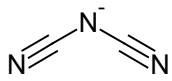
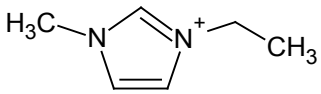
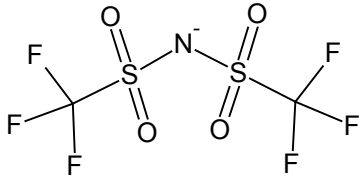
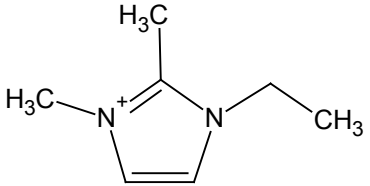
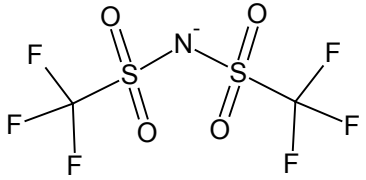
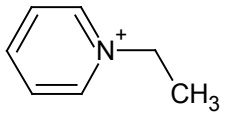
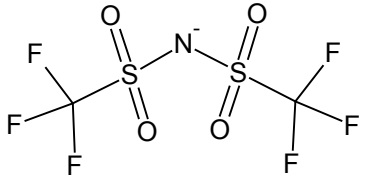
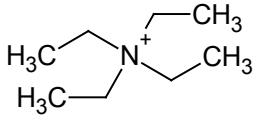
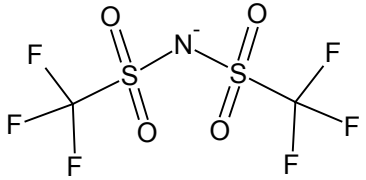
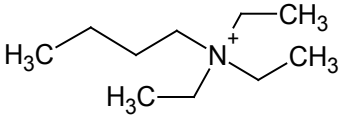
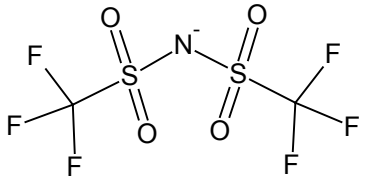
ID	Abbreviation	Structure & Name		Properties		
		Cation	Anion	μ @ 293K	ρ @ 298K	T_m
				Pa.s	Kg/m^3	K
21	$[m_2Im]mSO_4$			0.0928	1328	308.9
		1,3-dimethyl imidazolium	Methylsulfate			
22	$[emIm]CF_3SO_3$			0.0500	1384	262.2
		1-ethyl-3-methyl imidazolium	Triflate			
23	$[bmIm]oSO_4$			0.8745	1060	307.6
		1-butyl-3-methyl imidazolium	Octylsulfate			

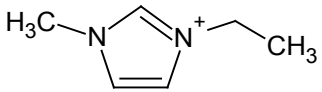
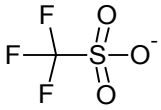
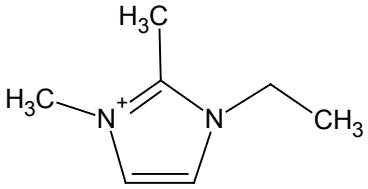
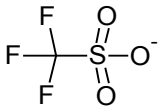
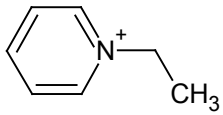
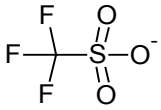
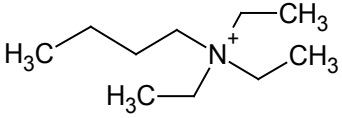
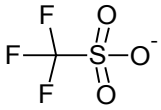
Table B.4: Candidate ionic liquid solutions enumerated from exhaustive search in latent property space.

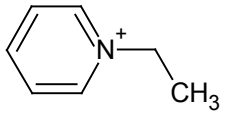
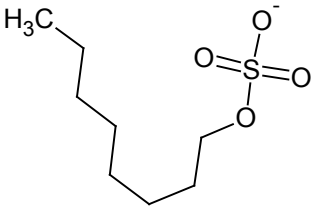
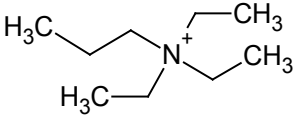
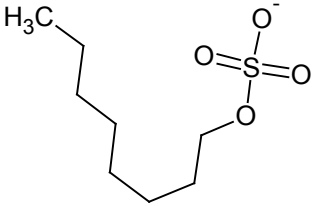
	Abbreviation	Structure & Name		Latent Properties		
		Cation	Anion	q_1	q_2	q_3
1	[meIm]mSO_4			0.0185	0.0045	-0.0293
		1-methyl-3-ethyl imidazolium	Methyl sulfate			
2	[ePy]mSO_4			0.0216	-0.0039	-0.0140
		1-ethylpyridinium	Methyl sulfate			
3	[e_3pNH_4]mSO_4			0.0457	0.0292	0.0147
		N,N,N- triethyl propanaminium	Methyl sulfate			
4	[meIm]PF_6			0.0113	0.0193	-0.0314
		1-methyl-3-ethyl imidazolium	Hexafluorophosphate			

	Abbreviation	Structure & Name		Latent Properties		
		Cation	Anion	q_1	q_2	q_3
5	$[ePy]PF_6$			0.0144	0.0109	-0.0160
		1-ethylpyridinium	Hexafluorophosphate			
6	$[e_3eNH_4]PF_6$			0.0384	0.0440	0.0126
		N,N,N- triethyl propanaminium	Hexafluorophosphate			
7	$[meIm]BF_4$			0.0111	0.0231	0.0012
		1-methyl-3-ethyl imidazolium	Tetrafluoroborate			
8	$[e2,3m_2Im]BF_4$			0.0050	0.0185	-0.0162
		1-ethyl-2,3-dimethyl imidazolium	Tetrafluoroborate			
9	$[ePy]BF_4$			0.0142	0.0148	0.0165
		1-ethylpyridinium	Tetrafluoroborate			

	Abbreviation	Structure & Name		Latent Properties		
		Cation	Anion	q_1	q_2	q_3
10	$[e_3eNH_4]BF_4$			0.0028	0.0128	-0.0302
		N,N,N- triethyl ethanaminium	Tetrafluoroborate			
11	$[e_3bNH_4]BF_4$			0.0290	0.0557	-0.0135
		N,N,N- triethyl butanaminium	Tetrafluoroborate			
12	$[e_3pNH_4]dCN$			0.0427	0.0068	-0.0130
		N,N,N- triethyl propanaminium	Dicyanamide			
13	$[meIm]Tf_2N$			0.0091	0.0069	-0.0002
		1-methyl-3-ethyl imidazolium	Bis(trifluromethylsulfonly)-amide			

	Abbreviation	Structure & Name		Latent Properties		
		Cation	Anion	q_1	q_2	q_3
14	$[e2,3m_2Im]Tf_2N$			0.0030	0.0022	-0.0176
		1-ethyl-2,3-dimethyl imidazolium	Bis(triflurомethylsulfonly)-amide			
15	$[ePy]Tf_2N$			0.0121	-0.0015	0.0151
		1-ethylpyridinium	Bis(triflurомethylsulfonly)-amide			
16	$[e_3eNH_4]Tf_2N$			0.0008	-0.0035	-0.0316
		N,N,N- triethyl ethanaminium	Bis(triflurомethylsulfonly)-amide			
17	$[e_3bNH_4]Tf_2N$			0.0270	0.0394	-0.0149
		N,N,N- triethyl butanaminium	Bis(triflurомethylsulfonly)-amide			

	Abbreviation	Structure & Name		Latent Properties		
		Cation	Anion	q_1	q_2	q_3
18	$[meIm]CF_3SO_3$	 1-methyl-3-ethyl imidazolium	 Triflate	0.0069	0.0182	-0.0118
19	$[e2,3m_2Im]CF_3SO_3$	 1-ethyl-2,3-dimethyl imidazolium	 Triflate	0.0008	0.0135	-0.0292
20	$[ePy]CF_3SO_3$	 1-ethylpyridinium	 Triflate	0.0099	0.0098	0.0035
21	$[e_3bNH_4]CF_3SO_3$	 N,N,N- triethyl butanaminium	 Triflate	0.0248	0.0507	-0.0265

	Abbreviation	Structure & Name		Latent Properties		
		Cation	Anion	q_1	q_2	q_3
22	[ePy]oSO$_4$	 1-ethylpyridinium	 Octylsulfate	0.0103	0.0116	-0.0241
23	[e$_3$pNH$_4$]oSO$_4$	 N,N,N- triethyl propanaminium	 Octylsulfate	0.0344	0.0446	0.0046