

**Integrated Multiscale Chemical Product Design using Property Clustering and
Decomposition Techniques in a Reverse Problem Formulation**

by

Charles C. Solvason

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
December 12, 2011

Keywords: Product Design, Chemical Product, Multiscale, Reverse Problem Formulation,
Property Clustering, Systems Engineering

Copyright 2011 by Charles C. Solvason

Approved by

Mario R. Eden, McMillan Associate Professor of Chemical Engineering
Ram B. Gupta, Woltosz Professor of Chemical Engineering
Virginia A. Davis, Associate Professor of Chemical Engineering
Nedret Billor, Associate Professor of Mathematics and Statistics

Abstract

In recent years, chemical engineers in the Process Systems Engineering (PSE) community have increasingly been using their skill set to solve problems in areas beyond the chemical manufacturing processes, focusing instead on the chemical products themselves. This trend reflects a movement within the field to meet the demands of an increasingly competitive and global consumer products marketplace that stresses being the first to market in conjunction with being the highest volume and lowest cost manufacturer. By definition, chemically formulated products deliver specific attributes to the consumer by manipulating a multitude of separate and often competing mechanisms in molecular architecture that operate over a wide range of length and time scales. Examples of chemical products include performance chemicals, paints, cosmetics, pharmaceuticals, proteins, semi-conductors, and foods, among others. Like process design, computer aided chemical product design is a complex programming problem. Generating, integrating, and managing the information, data, and knowledge at multiple length scales for use in various types of product design problems is a significant undertaking. The traditional approach to managing the complexity of this problem has been to compute information at smaller length scales and pass it to models at larger length scales by removing degrees of freedom (coarse-graining) with the objective being to predict macroscopic properties from molecular information. While often the most accurate method for predicting properties, this simulated approach has two limitations: (1) it has an immense computational cost due to hierarchical nesting and (2) it utilizes *a priori* knowledge of the molecular architecture (i.e. the number and types of atoms or electrons present). This dissertation covers the development of a

novel, alternative approach that allows for the simultaneous design of a chemical product's molecular architecture across multiple scales using a reverse problem formulation, property clustering, and decomposition techniques. The developed framework is specifically designed to utilize experimental data, parameters, and models since the effectiveness of a chemical product is most often determined by consumer attributes based on consumer preference tests.

In this work, three specific methodologies are developed. The first method, Attribute Computer Aided Mixture (Blend) Design (aCAM^bD) is an extension of Computer Aided Mixture (Blend) Design (CAM^bD) and includes experimental data and regression models, specifically, Scheffe canonical and Cox polynomial models. Necessary adjustments to the original clustering algorithm are identified and the design method is rewritten accordingly. The end result is a method capable of performing a mixture design on any chemical constituent data set across multiple length scales, as long as accurate attribute-component models can be established. A case study mixture design of spun yarn is presented to illustrate the method.

The second method developed is Attribute Computer Aided Molecular Design (aCAMD). It is an extension of Computer Aided Molecular Design (CAMD) to include experimental data and regression models, while continuing to use group contribution method (GCM) based property models. The technique uses design of experiments (DOE) to generate an attribute-property relationship and maps the attribute information into a property domain where molecular design can proceed. Adjustments to the property clustering algorithm are made to reflect the new design approach. The result is a method capable of performing molecular design on any attribute data set as long as a strong relationship between attribute and property models can be established. A case study involving the molecular design of environmentally benign refrigerants is presented to illustrate the method.

The third method developed is Characterization Based Computer Aided Molecular Design (cCAMD). It is a wholly new method that addresses the limitations of aCAM^bD and aCAMD, namely difficulty in finding suitable attribute-component and attribute-property models for complex chemical products. The method uses characterization tools like infrared and near-infrared spectroscopy (IR/NIR) to generate a set of data from a chemical constituent training set and then applies decomposition algorithms like principal component analysis (PCA) to find the underlying latent variable data structure. A parameterization of the data structure into a characterization based group contribution method (cGCM) follows. Attribute data is then mapped into the latent domain using a separate principal component regression (PCR) or partial linear regression on to latent surfaces (PLS) model. A molecular design is then performed in the latent domain. The resulting method is capable of performing structured molecular design across multiple scales for any system of attributes whose molecular architecture can be adequately described by characterization methods. A case study on the particle design of pharmaceutical excipients for an acetaminophen tablet is presented to illustrate the method.

Acknowledgments

I thank my advisor Dr. Mario Eden for introducing me to this research topic and providing an incredible opportunity to succeed. His advice, patience, and direction have made this job such a joy. I would like to thank my coworkers, Dr. Norman Sammons and Dr. Nishanth Chemmangattuvalappil for providing support during my research investigation. I would also like to acknowledge Dr. Jon Gabrielsson for providing his research data on excipients for direct compressed acetaminophen tablets. This data set was crucial to my proof of concept case study that tied this research together. I would like to thank Dr. Christopher Roberts for giving me the opportunity to return to the city of Auburn which has been an outstanding experience that I will never forget. And finally, I would like to acknowledge Dr. Virginia Davis, Dr. Ram Gupta and Dr. Rafiqul Gani for inspiring aspects of this research.

In my personal life I would also like to especially thank my wife, Lisa Trinh. Her encouragement to pursue my dreams is the reason I was able to return to Auburn University and conduct this research. I also thank my parents, Jim and Darlene Solvason, for instilling in me the principles of work and perseverance and my brother, Jonathan Solvason, for keeping up my spirits. Finally, to my grandmother Helen Word and my late grandfathers Brig. General Charles E. Word of the USAF (ret.) and John C. Solvason: Mission accomplished! I love you all.

Table of Contents

Abstract	ii
Acknowledgments	v
List of Figures	xii
List of Tables	xvii
Nomenclature	xxi
Chapter 1: Introduction to Process Systems Engineering	1
Chapter 2: Product Design for Chemical Engineers	11
2.1. An Emerging Paradigm	11
2.2. Types of Chemical Products	15
2.3. Chemical Product Scales	17
2.4. Other Definitions	21
2.5. General Design Approach	22
Chapter 3: Experimental Design of Chemical Products	32
3.1. Mixture Design using Experiments	33
3.2. Experimental Design using Multivariate Data	49
3.2.1. Characterization Techniques	51
3.2.2. Principal Component Analysis & Regression	61
3.2.3. Partial Least Squares on to Latent Surfaces (PLS)	66
3.2.4. Mixture Design and Optimization with Latent Variables	68

3.3. Summary	71
Chapter 4: Computer Aided Design of Chemical Products	72
4.1. Mathematical Formulation of Chemical Product Design	73
4.1.1. Searching a Database	75
4.1.2. Generating Candidate Solutions	78
4.2. Useful Property Prediction Techniques	81
4.2.1 Group Contribution Methods	86
4.3. Computer Aided Mixture Design (CAM ^b D).....	91
4.4. Computer Aided Molecular Design (CAMD)	92
4.5. Summary	94
Chapter 5: Motivation and Challenges	95
5.1. Current Limitations of Chemical Product Design	98
5.2. Reverse Problem Formulation	101
5.3. Challenges Addressed in this Dissertation.....	103
Chapter 6: Methods for Multiscale Chemical Product Design	110
6.1. Property Clustering Algorithms	110
6.2. Computer Aided Mixture Design (CAM ^b D) with Clusters	120
6.3. Computer Aided Molecular Design (CAMD) with Clusters	133
6.4. Summary	146
Chapter 7: Attribute-Computer Aided Mixture Design (aCAM ^b D)	148
7.1. Integrating Attribute Data & Models with Property Clustering	149
7.1.1. Canonical Models	159
7.1.2. Polynomial Models	160

7.2. Attribute Clustering Algorithms.	165
7.3. aCAM ^b D using Experimental Data and Data Driven Models.	168
7.4. Case Study: Polymer Spun Yarn Mixture Design	181
7.5. Summary	195
Chapter 8: Attribute-Computer Aided Molecular Design (aCAMD)	198
8.1. Integrating Attribute Data & GCM Models with the RPF	199
8.2. Attribute & Property Clustering Algorithms.	206
8.3. aCAMD using Experimental Data and GCM Models.	217
8.4. Case Study: Environmentally Benign Refrigerant Design	228
8.5. Summary	270
Chapter 9: Characterization based Computer Aided Molecular Design (cCAMD)	272
9.1. Integrating Multivariate Data & Models with the RPF	275
9.1.1. Characterizing the System with IR/NIR Spectroscopy.....	278
9.1.2. Developing the Latent Property Domain	294
9.2. Latent Property Clustering Algorithms.....	311
9.3. cCAMD using Characterization Data and cGCM Models.	325
9.4. Case Study: Excipient Design for Direct Compressed Tablets	330
9.5. Summary	344
Chapter 10: Conclusions	346
10.1. Achievements.....	347
10.2. Challenges and Future Work	349
References	351
Appendix A1: Additional Chemical Product Design Techniques	375

A1.1. Multi-Criteria Decision Making (MCDM)	375
A1.2. Ontology.....	383
A1.3. Mathematical Optimization.....	385
Appendix A2: Additional Characterization Techniques	394
A2.1. Ultraviolet-Visible Spectroscopy.....	394
A2.2. Nuclear Magnetic Resonance Spectroscopy	395
A2.3. Mass Spectrometry.....	397
A2.4. Chromatography (Gas, Liquid, or Gel Permeation).....	397
A2.5. Scattering and Diffraction Techniques (Light, Neutron, X-ray).....	398
A2.6. Other Characterization Techniques.....	401
Appendix A3: Additional Decomposition Tools	404
A3.1. Network Component Analysis	404
Appendix A4: Additional Physico-Chemical Property Models	407
A4.1. Simulation Techniques.....	407
A4.2. Topological Indices	414
Appendix A5: Acetaminophen Case Study Supporting Data	422
Appendix A6: IR & NIR Molecular Group Absorbance Frequencies	423
A6.1. Methine CH.....	423
A6.2. Methylene CH ₂	423
A6.3. Hydroxyl OH (w/o H-bond).....	424
A6.4. Hydroxyl OH (w/ H-bond).....	424
A6.5. Secondary Alcohol CHOH (w/ H-bond).....	424
A6.6. Primary Alcohol CH ₂ OH (w/ H-bond)	425

A6.7. Alcohol CHCH_2OH (w/ H-bond).....	425
A6.8. Aliphatic Ether O	426
A6.9. Ether CH_2O	426
A6.10. Aldehyde HCO (w/o H-bond).....	427
A6.11. Aldehyde HCO (w/ H-bond).....	427
A6.12. Saturated Aliphatic Aldehyde CH_2CHO (w/o H-bond).....	427
A6.14. Vinyl CH_2CH	429
A6.15. Vinylidene CH_2C	429
A6.16. <i>cis</i> -Vinylene CHCH	430
A6.17. <i>trans</i> -Vinylene CHCH	430
A6.18. Aliphatic Methyl CH_3	430
A6.19. Aryl Methyl CH_3	431
A6.20. Aliphatic Methoxy OCH_3	431
A6.21. Aryl Methoxy OCH_3	432
A6.22. Aryl OH (w/o H-bond).....	433
A6.23. Aryl OH (w/ H-bond).....	433
A6.24. Tetramethyl $\text{C}(\text{CH}_3)_3$	433
A6.25. Alkyl Peroxide OO	434
A6.26. Saturated Aliphatic Ester COO	434
A6.27. Saturated Aliphatic Methyl Ester COOCH_3	434
A6.28. Saturated Aliphatic Ethyl Ester $\text{COOCH}_2\text{CH}_3$	435
A6.29. Acrylate Ester CH_2CHCOO	436
A6.30. Methacrylate Ester $\text{CH}_2\text{C}(\text{CH}_3)\text{COO}$	437

A6.31 Other Groups.....	438

List of Figures

Figure 2.1. Multi-length Scale Biochemistry Approach to Product Design	18
Figure 2.2. An Example of a Multi-length and Time Scales for a Powdered Detergent	19
Figure 2.3. Multi-length and Time Scales Associated with Computational Simulation	20
Figure 2.4. Multi-scale Chemical Product Supply Chain	21
Figure 2.5. The Stage-Gate Development Process for Chemical Product Design.....	28
Figure 3.1. Mixture Design Point and Response Surface Simplex Diagram.....	38
Figure 3.2. Parameterization of Component Design Space in the Simplex Diagram	42
Figure 3.3. Overview of Characterization Techniques	52
Figure 3.4. The Process-Structure-Property Relationship for Alloy Steel	53
Figure 3.5. WAXS Characterization Structure of High Performance Alloy Steel	53
Figure 3.6. IR Spectra of Mannitol	56
Figure 3.7. NIR Spectra of Mannitol	58
Figure 3.8. Graphical Representation of PCA	63
Figure 3.9. Scree Plot of a Covariance-PCA IR and NIR Spectroscopy Data	64
Figure 3.10. Graphical Representation of PLS	67
Figure 4.1. Methods to find Property Models and Parameters	84
Figure 4.2. Property Model Development Scheme	85
Figure 5.1. Computational Approach to Multiscale Property Estimation	97
Figure 5.2. Streamlined Multiscale Property Estimation.....	99

Figure 5.3. Chemical Product Design using Consumer Attributes	100
Figure 5.4. Reverse Problem Formulation of Chemical Product Design	102
Figure 5.5. Reverse Problem Formulation for Multiscale Chemical Product Design	104
Figure 6.1. Property Clustering Binary Mixing Diagram	114
Figure 6.2. CAM ^b D Property Clustering Conversion Algorithm	117
Figure 6.3. CAMD Property Clustering Conversion Algorithm	119
Figure 6.4. Property Clustering Ternary Mixing Algorithm.....	126
Figure 6.5. CAM ^b D Candidate Generation Algorithm	130
Figure 6.6. CAM ^b D test on Pure Components.....	131
Figure 6.7. CAM ^b D test on Binary Mixtures.....	132
Figure 6.8. CAM ^b D test on Ternary+ Mixtures.....	133
Figure 6.9. GCM in Property Clusters	136
Figure 6.10. CAMD Candidate Generation Algorithm	142
Figure 6.11. CAMD test on Pure Group Molecules	143
Figure 6.12. CAMD test on Binary Group Molecules.....	144
Figure 6.13. CAMD test on Ternary+ Mixtures	145
Figure 7.1. Ternary Property Cluster Diagram of a Six Component Mixture	153
Figure 7.2. Negative Property Clusters in an Eight Component Mixture.....	154
Figure 7.3. Negative Property Cluster Domains.....	156
Figure 7.4. Property Cluster Reference Comparison for a Three Component Mixture.....	158
Figure 7.5. aCAM ^b D Property Clustering Conversion Algorithm	166
Figure 7.6. Property Model Comparison Cluster Diagram.....	172
Figure 7.7. aCAM ^b D with Cluster Generation Algorithm.....	173

Figure 7.8. aCAM ^b D test on Pure Components	174
Figure 7.9. aCAM ^b D test on Binary Mixtures	175
Figure 7.10. aCAM ^b D test on Ternary+ Mixtures	176
Figure 7.11. Conventional Mixture Diagrams for a Mixture of Spun Yarn	184
Figure 7.12. Property Clustering Diagram of Canonical Models of Spun Yarn.....	187
Figure 7.13. Property Clustering Diagram of Polynomial Models of Spun Yarn	190
Figure 7.14. Model Feasibility Region in Property Cluster Space	194
Figure 8.1. Reverse Problem Formulation Problem for Consumer Attributes	202
Figure 8.2. Overlapping of Groups	211
Figure 8.3. aCAMD Property Clustering Conversion Algorithm.....	215
Figure 8.4. aCAMD Candidate Generation Algorithm.....	225
Figure 8.5. aCAMD test on Pure Group Molecules.....	226
Figure 8.6. aCAMD test on Binary+ Group Molecules.....	227
Figure 8.7. Attribute Cluster Diagram of Refrigerant Design	236
Figure 8.8. Predicted vs. Actual Attribute Plots	241
Figure 8.9. Property Cluster Diagram of FR and MFR	253
Figure 8.10. Property Cluster Diagram of pFR and pMFR	254
Figure 8.11. Property Cluster Diagram of Additive FR.....	256
Figure 8.12. Property Cluster Diagram of Additive Groups.....	264
Figure 8.13. Property Cluster Diagram of Additive Candidate Molecules.....	266
Figure 8.14. Predicted vs. Actual H_v Plots.....	267
Figure 8.15. Predicted vs. Actual T_c and P_c Plots	268
Figure 9.1. Reverse Problem Formulation for Multiple Scales	276

Figure 9.2. Rotated Reverse Problem Formulation for Multiple Scales.....	276
Figure 9.3. cGCM Construction Example Highlighting Isobutanol	290
Figure 9.4. Raw NIR Absorbances of L-Glutamic Acid	293
Figure 9.5. IR Transmittance and NIR Absorbance of Mannitol	300
Figure 9.6. p-Standardized and Centered IR and NIR Responses of the Filler Training Set ...	301
Figure 9.7. Gamma Q-Q plot of filler excipients.....	301
Figure 9.8. n,p-Standardized and Centered IR / NIR Responses of the Filler Training Set	306
Figure 9.9. Latent Variable Predicted IR and NIR Responses of Mannitol	308
Figure 9.10. PCA Loading of the Filler Training Set	315
Figure 9.11. Clustered Loadings of the Filler Training Set	316
Figure 9.12. PCA Scores of the Filler Training Set.....	319
Figure 9.13. Clustered Scores of the Filler Training Set	320
Figure 9.14. IR Transmittance Methylene Group Response Plot	321
Figure 9.15. cCAMD Cluster Conversion Algorithm.....	324
Figure 9.16. cCAMD Candidate Generation Algorithm.....	328
Figure 9.17. Scree Plot of Proof of Concept.....	335
Figure 9.18. cCAMD Training Set	343
Figure A1.1. Decision Analysis Tree.....	382
Figure A1.2. Ontological Relationships for a Pharmaceutical Process	384
Figure A1.3. Branch-and-Bound Algorithm Example.....	387
Figure A1.4. Branch-and-Bound Algorithm Work Flow.....	388
Figure A1.5. Flowsheet of MINLP Optimization.....	390
Figure A1.6. Genetic Algorithm Work Flow.....	393

Figure A3.1. Network Component Analysis.....	406

List of Tables

Table 2.1. Chemical Product Types.....	15
Table 3.1. Relative Band Intensities for C-H Stretch	59
Table 3.2. Group Structures with First Order Coupling	60
Table 4.1. Designed chemical structures for the partition coefficient	77
Table 6.1. Common Physical-Chemical Properties used in Chemical Product Design.....	112
Table 6.2. CAM ^b D Property Clustering Conversion Algorithm.....	116
Table 6.3. CAMD Property Clustering Conversion Algorithm.....	118
Table 6.4. CAM ^b D Candidate Generation Algorithm using Clusters.....	128
Table 6.5. CAMD Candidate Generation Algorithm using Clusters	140
Table 7.1. aCAM ^b D Property Clustering Conversion Algorithm for Canonical Models.....	166
Table 7.2. aCAM ^b D Property Clustering Conversion Algorithm for Polynomial Models.....	167
Table 7.3. aCAM ^b D Candidate Generation Algorithm using Clusters.....	177
Table 7.4. Simplex Lattice Design MDOE for Polymer Spun Yarn	181
Table 7.5. Scheffe and Cox Model Regression Coefficients	182
Table 7.6. Nondimensionalized Property Operators and References	185
Table 7.7. Property Clusters and <i>AUP</i> of Canonical Property Operators.....	186
Table 7.8. Property Clusters and <i>AUP</i> of Experimental Mixtures.....	186
Table 7.9. Property Clusters and <i>AUP</i> of Feasibility Region	187
Table 7.10. Property Clusters and <i>AUP</i> of Standard Reference Property Operators.....	188

Table 7.11. Property Clusters and <i>AUP</i> of Polynomial Property Operators.....	188
Table 7.12. Property Clusters and <i>AUP</i> of Pseudo Feasibility Region.....	189
Table 7.13. Binary Mixtures that meet Rule 1 and 2	192
Table 7.14. Ternary Mixtures that meet Rule 1, 2, and 3	192
Table 8.1. aCAMD Cluster Conversion Procedure	215
Table 8.2. aCAMD Candidate Generation Procedure	223
Table 8.4. List of Potential Refrigerants for the Case Study	230
Table 8.5. MDOE for the Case Study	232
Table 8.6. MDOE Attribute Responses	234
Table 8.7. Calculated Property Responses.....	238
Table 8.8. Informative Statistics for the Attribute-Property Relationships	239
Table 8.9. Informative Statistics for the Property-Attribute Relationships	243
Table 8.10. Bounds on the FR	243
Table 8.11. Bounds on the pFR	244
Table 8.12. Bounds on the MFR.....	244
Table 8.13. Bounds on the pMFR.....	245
Table 8.14. Pure Component Effects	245
Table 8.15. Nondimensional Property Operators of the Training Set	246
Table 8.16. Property Clusters and <i>AUP</i> of the Training Set.....	248
Table 8.17. Nondimensional Property Operators of the Pure Component Effects	249
Table 8.18. Clusters and <i>AUP</i> of the Pure Component Effects	250
Table 8.19. Nondimensional Property Operators of the FR, MFR, pFR, and pMFR.....	250
Table 8.20. Property Cluster bounds on the FR.....	251

Table 8.21. Property Cluster Bounds on the MFR.....	251
Table 8.22. Property Cluster Bounds on the pFR	252
Table 8.23. Property Cluster Bounds on the pMFR	252
Table 8.24. Clusters and <i>AUP</i> of the Additive Domain.....	256
Table 8.25. Nondimensional Property Operators of the Additive FR	257
Table 8.26. Property Data of First Order Groups	258
Table 8.27. Nondimensional Property Operators of First Order Groups.....	259
Table 8.28. Maximum No. First Order Groups	260
Table 8.29. Properties of 2 nd Order Groups	261
Table 8.30. Nondimensional Property Operators of Second Order Groups	262
Table 8.31. First Order Groups Clusters.....	263
Table 8.32. Candidate Molecules.....	265
Table 9.1. Fundamental Vibrations.....	281
Table 9.2. Approximation of Transmittance IR/NIR Active Tones	282
Table 9.3. Derived IR Absorption of Methylene Group	285
Table 9.4. Filler Excipients.....	297
Table 9.5. SIMCA-P PCA Scores for Filler Excipients.....	303
Table 9.6. Matlab and JMP PCA Scores for Filler Excipients	305
Table 9.7. CAMD Candidate Generation with Latent Variables.....	329
Table 9.8. CAMD Candidate Excipients	330
Table 9.9. Property Descriptors	332
Table 9.10. Latent Variable Scores.....	333
Table 9.11. Latent Variable Loadings.....	335

Table 9.12. Design or Experiments.....	337
Table 9.13. Regression Coefficients for Attribute-Latent Property Relationship	338
Table 9.14. Target Attributes	338
Table 9.15. Latent Variable Targets	339
Table 9.16. Predicted Property Descriptor Targets.....	339
Table 9.17. Standardized Latent Variable Targets.....	339
Table 9.18. Groups for Molecular Design of Excipient.....	341
Table 9.19. Designed Molecules.....	344
Table A1.1. MCDM Needs and Wants for PDK Inhibitor	377
Table A1.2. Scoring System for PDK Inhibitor Example	379
Table A1.3. MCDM Ranking for PDK Inhibitor Example	381

Nomenclature

Latin Symbols

a	total number of product attributes	C^L	cluster of loadings
A	array of consumer attributes	C^Q	cluster of scores
A_L	linear algebra constant for x -axis	C_M^s	standard reference cluster of a mixture
AUP	augmented property index	C^z	pseudo cluster of a component
AUP_M	augmented property index of mixture	C_M^z	pseudo cluster of a mixture
AUP_M^s	standard mixture AUP	d	vector of structural descriptors
AUP_M^z	pseudo mixture AUP	D	total number of expressions
AUP_F	feasibility region AUP	e	design point or experiment identifier
b	latent property target identifier	E	residual variable array
B	array of regressors	E_Y	residual of response Y
B_L	linear algebra constant for y -axis	E_P	residual of response P
c	concentration in spectrophotometer	f_1	first order group subset identifier
C	cluster of component properties	f^{abs}	absorbance probability density function
C_M	cluster of mixture properties	F	total number of first order groups
C^M	cluster of molecular group properties	F_c	total number of group combinations
C_i^M	cluster of molecule properties	$F(\xi)$	F distribution critical value
C^U	feasibility upper limit cluster	FBN	free bond number of molecular structure
C^L	feasibility lower limit cluster	F_{OBJ}	objective function for optimization

g	group identifier (of any order)	K_L	linear algebra intercept constant
g_1	first order group identifier	K_L^{ref}	linear algebra reference constant
g_2	second order group identifier	l	total number of latent properties
g_3	third order group identifier	l_b	beam path length in spectrophotometer
G	total number of second order groups	l_1	vector of compositional lower limits
G_S	total number of overlapping groups	l_2	vector of architectural lower limits
G_f	standard Gibbs free energy	l_3	vector of combinatorial lower limits
h	Planck's constant	L	array of latent component loadings
h_1	vector of composition constraints	m	total number of latent variables
h_2	vector of process constraints	M	atomic mass
h_3	vector of property model constraints	M_X	maximum AUP
H	UNIUQAC surface area parameter	n	total number of experiments
H_f	standard enthalpy of formation	n_g	total number of groups (or any order)
H_v	standard enthalpy of vaporization	n_q	quantum number vibrational integer
H_{fus}	standard enthalpy of fusion	N	total number of atoms
i	chemical constituent identifier	N_v	number of vibrational motions
I	attenuated energy in a spectrophotometer	N_{RG}	total number of rings in structure
I_o	incident energy in a spectrophotometer	N_R	total number negative cluster regions
j	physical-chemical property identifier	N_T	number of types of cluster regions
J	UNIUQAC volume parameter	p	total number of properties
k	chemical product attribute identifier	P	array of physical-chemical properties
K	bond force constant	P_c	critical pressure
K_{OW}	octanol-water coefficient	P_M	array of property mixtures

P^{abs}	spectroscopic absorbance	T^2	Hoetelling's T^2 distribution statistic
P^T	spectroscopic transmittance	u	total number of chemical constituents
q	latent property identifier	u_1	vector of compositional upper limits
q_i	relative molecular surface area	u_2	vector of architectural upper limits
Q^2	predicted fit of an empirical model	u_3	vector of combinatorial upper limits
Q	group contribution to vdW surface area	U	array of latent property scores
r	molecular van der Waals volume	v_1	dissimilar first order group identifier
R^2	fitness of an empirical model	V_c	critical volume
R	group contribution to vdW volume	w	binary dissimilar constituent identifier
s	standard reference mixture	W	array of latent property loadings
s^2	variance	w_S	second order group factor
S	tensor of variance and covariance	w_T	third order group factor
S	summation of latent property loadings	x	vector of chemical compositions
S_{tot}	total sum of squares	x_i	mass fraction of component i
SS_{res}	residual sum of squares	x_a	anharmonicity shift
SS_{PRE}	predicted residual sum of squares	X	array of chemical constituent fractions
t	total number of latent property targets	X_{CC}	x-axis value in Cartesian coordinates
T	array of latent component scores	y	independent response variable (A or P)
T_m	normal melting point	$\bar{y}$	mean of independent response variable
T_m	normal boiling point	Y	array of attribute or property responses
T_c	critical temperature	Y_{CC}	y-axis value in Cartesian coordinates
Th	total number of third order groups	z	design model identifier
T_M	array of latent mixture scores	z^M	total number of models in design

Nomenclature

Greek Symbols

α	total number of atoms	λ^{bar}	peak wavelength
β_o	polynomial reference regressor	η_f	number of overlapped 2 nd order groups
β_i	polynomial regressor of component i	η_s	number of overlapped 2 nd order groups
β_{ii}	nonlinear polynomial regressor of i	μ	mixture level identifier
β_{iw}	polynomial regressor of interaction iw	μ_g	group mixture level identifier
β_i^*	canonical regressor of component i	μ^{max}	maximum mixture type
β_{iw}^*	canonical regressor of interaction iw	ρ	density of component
δ	cluster mixing fraction	ρ_M	density of mixture
E_θ	band energy	σ	degree of fractionation
ε	absorptivity	τ	atom identifier
ξ	probability	θ	integer of vibrational quantum number
ϕ	number of factors in a design	ν	vibrational frequency
γ_i^c	pure species activity coefficient	ν^o	vibrational overtone frequency
γ	ternary dissimilar constituent identifier	Ω	normalized property operator
Γ	total number of responses (a or p)	Ω_M	normalized property operator of mixture
κ	linear constant	Ω^{ef}	normalized reference property operator
λ	wavelength	Ω_M^s	normalized property operator of standard

Ω_M^z normalized operator of pseudo mixture

ψ property operator

ψ^M property operator of mixture

ψ^{ref} reference property operator

Chapter 1

Introduction to Process Systems Engineering

The focus of this dissertation is on the development of novel process system engineering (PSE) techniques for the design of chemical products having molecular architectures that exist in multiple length scales. PSE is generally defined as the research and development of systematic procedures for the design and operation of chemical process systems, from the molecular length scale through the supply-chain mega-length scale (Grossman and Westerberg 2000). It was first described in a Special Volume of the AIChE Symposium Series in 1961 and is primarily focused on the design and optimization of macro-scale process systems (Grossman and Westerberg 2000). The heart of PSE research is the discovery of concepts and models for the computationally efficient prediction of the performance of a system of unit operations. Unit operations (e.g. fluid flow, heat and mass transfer, thermodynamic phase behavior, reactions, etc.) are basic steps in a process described by universal physical laws and may be used across all chemical industries. It is common for a single process to contain multiple processing steps, and each processing step to include multiple unit operations, resulting in a highly complex system. As a result, PSE solution techniques demand the use of software simulators (e.g. AspenONE®, ProMax®, etc.) and/or mathematical solvers (e.g. GAMS®, MatLab®, etc.). In the chemical engineering curriculum, PSE concepts are usually introduced as a senior capstone course that pulls together the unit operation topics covered in earlier courses, making it a cornerstone of chemical engineering.

Recently, many of the industries that employ chemical engineers have experienced increased global competition which has placed considerable pressure on their profit margins. For a number of companies it has become apparent that it is no longer sustainable to manufacture only commodity chemicals, defined as chemicals produced in bulk, having minimum variation between manufacturers and simple molecular architecture. Traditionally, because of the uniformity of commodity chemical products, any gains in profit margin were due to linear improvements in process configurations and production volume. However, with the increase in global competition, linear improvements in process design are not quick enough or large enough to hold on to market share. As a result, successful manufacturers are making a nonlinear shift in focus from process optimization to *product innovation* by diversifying their product portfolios to include more specialty chemical products. Specialty chemicals are produced in small quantities, having simple or complex molecular architecture and large variation between manufacturers. Examples of specialty chemicals include pharmaceutical ingredients, biocides, intermediates, and many other types of specialty market products. By shifting focus to manufacture specialty chemical products for niche markets, manufacturers can dominate market share, drive up prices for their products, and offset pressure from global competition. Further analysis of the emergence of specialty chemical products is provided in Section 2.1 and Section 2.2.

One of the interesting aspects of the shift toward specialty chemical products is that the relationship between the molecular architecture of a specialty chemical and its physical and chemical properties is often more complex than for a commodity chemical. Molecular architecture, by definition, includes information related to atoms and bonds at the molecular length scale as well as nanoscale, microscale, mesoscale, and macroscale structure information related to aspect ratio, crystallinity, morphology, and many more orientation specific

arrangements of chemical structure. A discussion of the length scales and their relationship to specific molecular architectures of interest to chemical engineers is presented in Section 2.3.

Since the unique combination of the molecular architectural arrangements is often what gives a specialty chemical its unique set of properties, developing a thorough understanding of this relationship is vital. Research in this area has traditionally been performed by the natural sciences, but a compressed time-to-market schedule has brought it under the domain of chemical engineers who seek to integrate it with conventional chemical engineering concepts. As a result, research in chemical engineering has broadened to include the quantification of the relationships between physical-chemical properties and molecular architecture. Simultaneously, the scope of research in the PSE community has broadened to include the development of design tools meant to utilize newly developed length dependent structure-property relationships. 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. Instead, specialty chemicals rely on a set of customer needs or consumer attributes to define the chemical product. Developing techniques to discover and optimize the molecular architecture that delivers these attributes is the focus of research in the PSE community and is discussed in more detail in Section 2.4.

Depending on the nature of the molecular architecture, physical-chemical properties, and/or consumer attributes, the design techniques employed can be largely experimental or computational in nature. Experimental techniques are preferred when the molecular architecture-property relationship is poorly understood, the solution is in a validation stage, or high

computational costs preclude the use of computer aided methods. More detail on experimental design techniques and their development in the PSE community is presented in Chapter 3.

Alternatively, computational design techniques are preferred when the molecular-architecture-property relationship is well defined, the design is in an exploratory stage, or computational costs are reasonable. The PSE developed mathematical formulation of this problem is presented in Section 4.1 and some useful property models are discussed in Section 4.2. Two recently developed computational design techniques, computer aided mixture design (CAM^bD) and computer aided molecular design (CAMD), are presented in Section 4.3 and Section 4.4, respectively.

Beyond the two design approaches detailed in Chapter 3 and Chapter 4, PSE approaches to product design and development have proceeded in one of three directions: (1) the application of macroscale systems techniques to other length scales, (2) the development of new techniques specific to non-macroscale domains, and (3) the development of tools capable of communicating molecular architecture and property information between scales. Many authors have conducted research in the extension of PSE techniques to other length scales, such as the work on the supply chain at the megascale by Grossman and Westerberg (2000) and the extension of the conventional PSE formulation to the molecular scale by Gani (2004). Others have focused on developing techniques specific to a particular scale, such as the prediction of crystalline structure (Karamertzanis *et al.* 2009; Doherty and Green 2004) and protein structure (McAllister and Floudas 2008) at the microscale. The third area involving the development of multiscale design methods has only recently come under focus of the PSE community and it is in this area that a significant contribution has been made in this dissertation.

In Chapter 5, the motivation for the development of a PSE technique to aid in the design of products with molecular architecture existing at multiple scales is presented. Most authors, such as Fermeglia and Prici (2009), have approached this multiscale design problem in a similar manner as the natural sciences, computing information at smaller length scales and passing it to models at larger length scales by removing degrees of freedom (coarse-graining). While often the most accurate method for predicting properties, this simulated approach has two limitations: (1) it has an immense computational cost due to hierarchical nesting and (2) it utilizes *a priori* knowledge of the molecular architecture (i.e. the number and types of atoms or electrons present). The large computational cost typically prevents an accurate modeling of mesoscopic structure such as the morphology of polymers without the use of constraints that significantly limit the degrees of freedom in the simulation (Fermeglia and Prici 2009). Furthermore, when this method is integrated within the product-process design framework, the computational intensiveness exponentially increases yet again since each projected molecular architecture must be simulated to determine its physical-chemical properties (Hill 2004). To minimize the computational cost in these types of problems, the property prediction simulations are typically approximated with constitutive equations based on structure descriptor models such as property based group contribution (GCM), quantitative structure property/activity relationships (QSPR/QSAR) using topological indices (TI), or chemometric based models. The introduction of structure descriptor models improves computational efficiency by avoiding nested simulation at the cost of prediction accuracy. However, these structure descriptor models are still computationally expensive when utilized in design algorithms because of their often highly nonlinear nature. Finding a method that addresses these limitations is a challenge and is discussed in more detail in Chapter 5.

Chapter 6 introduces a framework that addresses the multi-scale chemical product design challenges illustrated in Chapter 5. In particular, Gani (2004) and Eden *et al.* (2004) have shown that the reverse problem formulation (RPF) can be used to circumvent computational limitations by decoupling the constitutive equations from the constraint and balance equations to identify the property solution domain without committing to any components *a priori* (Gani and Pistikopoulos 2002; Eden *et al.* 2003; Gani 2004). The improvement in computational efficiency is achieved by only solving the systems of equations for chemical products with molecular architecture that reside in this property solution domain and is discussed in more detail in Section 6.1.

In addition to the reverse problem formulation, the framework proposed in this dissertation also makes use of the property clustering algorithm. The property clustering algorithm transforms physical-chemical properties into conserved surrogate property clusters that can be described by linear mixing rules, even if the operators themselves are nonlinear. At the heart of the algorithm is the use of transformation functions to linearize the expression for the physical-chemical property – molecular architecture relationship so that the duality of linear programming can be applied to solve the design problem in the lower dimensional property domain instead of the high dimensional molecular architecture space. A detailed discussion on how the cluster transformation algorithm works is presented in Section 6.2. Algorithms for conducting CAM^bD and CAMD with property clusters are presented in Sections 6.3 and 6.4, respectively.

One of the most important aspects in the design of specialty chemical products is the recognition that the effectiveness of a chemical product is determined by how well it meets the needs of the customer or consumer, not by molecular architecture or physical or chemical

properties. These needs are most often parameterized as consumer attributes and quantified using consumer preference experiments. As a result, consumer attributes tend to be poorly defined in terms of physical and chemical properties, making the design of chemical products using conventional approaches a difficult task. To alleviate this issue, this dissertation utilizes experimental data, parameters, and models in the design of specialty chemical products with molecular architecture that exists at multiple scales. The decision to use experimental data does not, however, preclude the use of more common physical and chemical properties. While the use of data places constraints on the application range of the solutions, it does not prevent the developed framework from being applied to systems described by solely deterministic relationships discussed in Appendix A4. Further discussion of data driven design of chemical products is presented in Chapter 6.

Three methods to address the current limitations of multiscale chemical product design of specialty chemicals based on consumer attributes are presented in this dissertation: (1) attribute based computer aided mixture (blend) design (aCAM^bD), (2) attribute computer aided molecular design (aCAMD), and (3) characterization based computer aided molecular design (cCAMD). Each of the developed methods utilizes the reverse problem formulation, property clustering, and consumer attribute data. Chapter 7 discusses the development of the aCAM^bD method to address situations where the attributes of interest are not conventional physical-chemical properties and mixtures of chemical constituents with different molecular architectures will meet the requirements of the chemical product. A case study on spun yarn of polymer filaments for marine applications is presented to illustrate the method.

In situations where mixture design will not suffice, but attributes remain the drivers of product performance, molecular design methodologies are utilized. Chapter 8 presents an

aCAMD method that first converts attribute data to traditional physical-chemical properties described by well-defined molecular architecture models, and then generates candidate solutions using a cluster-based molecular design technique. The benefit of this approach is that chemical products not present in the original training data set can be generated so long as they contain the same, underlying molecular architecture as the training data set. A case study design for an environmentally benign refrigerant is used to illustrate the method.

In the previously developed aCAMD method, attribute data was converted to physical-chemical properties described by group contribution method (GCM) based property models that ensured easy computation. However, GCM property models are limited to describing small molecules (< 30 atoms) with structural isomerism. Chemmangattuvalappil *et al.* (2010) developed a method for more complex molecular architectures using molecular signatures to convert QSPR/QSAR models based on topological indices into group based parameters. However, the method is constrained by the limitations of the QSPR/QSAR models, which typically only describe molecular length scales (< 60 atoms) and occasionally macromolecular length scales (< 60 atoms in the repeat unit of a polymer). QSPR/QSAR models describing physical-chemical property – molecular architecture relationships at other length scales or at multiple length scales are usually unavailable since the physics in these domains are still being investigated. As a result, researchers designing chemical products at these scales have relied on data-driven approaches based on chemical characterization.

Chemical characterization consists of a variety of techniques using spectroscopic, diffraction, and/or thermophysical techniques to determine the molecular architecture of a chemical product. One of the most common techniques is infrared spectroscopy (IR), which is capable of describing molecular architecture at multiple length scales. In Chapter 9, a

characterization based computer aided molecular design (cCAMD) framework is presented. Using a chemical product training set, IR/NIR spectroscopic data on the molecular architecture is first captured, and then converted to underlying latent variables using decomposition techniques like principal component analysis (PCA). Next, an experimental design is conducted to determine the effect of the latent variable structure on the consumer attributes of interest. Target consumer attributes are then mapped down to the latent variable domain and mixture design of the original training set is conducted. The same latent variable structure can be used to find new chemical products that were not part of the original training set by using a method developed in this dissertation called characterization based computer aided molecular design (cCAMD). First, recognizing that the absorbance measured by IR/NIR spectroscopy are based on group theory interactions with the incident radiation, group contributions based on characterization data are estimated using a newly developed characterization based group contribution method (cGCM). Characterization based groups are then converted to the same latent variable basis as the original training set and recombined to find new chemical products. The primary benefit of this approach is that the individual group contributions incorporate information from multiple length scales, decoupling the length scales, and circumventing the need for nested simulation typically found in multiscale design. A case study on the design of pharmaceutical excipients for direct compressed acetaminophen tablets using data provided by Gabrielsson *et al.* (2003) is presented in Chapter 9 to highlight the method.

This dissertation concludes with Chapter 10, where the achievements of this dissertation are summarized and future challenges are discussed. Special attention is paid to the contributions made by this author to the computer aided, data-driven design of chemical products using consumer attributes and characterization based approaches. Additionally, it should be

noted that the aCAM^bD approach described in Chapter 7 has been published in *Industrial and Engineering Chemistry Research* (Solvason *et al.* 2008), the aCAMD approach described in Chapter 8 has been published in *Computers and Chemical Engineering* (Solvason *et al.* 2008), and aspects of the cCAMD approach described in Chapter 9 has been published in parts in *Computer Aided Chemical Engineering* (Solvason *et al.* 2008, Solvason *et al.* 2009, Solvason *et al.* 2010). Other case studies performed by the author, but not presented in this dissertation, include the design of product pathways for the forest bio-refinery, published in *Integrated Biorefineries: Design, Analysis, and Optimization* (Batsy *et al.* 2011) and the design of self-assembling monolayers for chemical deposition.

Chapter 2

Product Design for Chemical Engineers

Product design, which is defined as the conversion of a conceptual idea into a tangible, manufactured object, is an important function of today's corporations. Increasing competition, fluctuating markets, and global supply chains have all worked to place high importance on delivering differentiated products to the consumer. Although consumer driven product design has always been appreciated as a major issue in the mechanical and electrical engineering sectors, in the chemical, pharmaceutical, and materials industries the systematic and efficient design of new products is a recent concern (Costa *et al.* 2006). Furthermore, consumer attributes are being pushed further up the development pipeline to the formulation scientists who are ill-equipped at utilizing this new source of information. Compounding the issue are the unique features of chemically formulated products that control the product's properties, namely the mechanisms that operate at a multitude of nested length scales. The effective utilization of this immense data structure to design chemical products is the subject of cutting edge research in the process systems engineering field and often referred to as an emerging paradigm within chemical engineering (Costa et 2006).

2.1. An Emerging Paradigm

“Unit Operations” is considered the first and most recognized paradigm in chemical engineering. Beginning in 1915, the concept of unit operations transformed the chemical

engineering discipline from a study of how to manufacture a specific commodity, to the systematic design of a sequence of unit operations that is applicable across all commodities (Wei, 1996). Although extremely useful, the unit operations paradigm proved inadequate for solving problems with increased molecular complexity (Hill 2009). As a result, a second paradigm involving the study of chemical science arose in 1950, best exemplified by Bird, Stewart, and Lightfoot's textbook on Transport Phenomena (Bird *et al.* 2002). It is focused on developing universal mathematical systems to describe the underlying, continuum based, chemical and physical sciences in order to solve relevant problems of interest (Hill 2009). This is the paradigm under which most chemical engineering research is conducted today and, although the discipline has moved into the biological sciences realm and new systems of interest have been incorporated, the mindset has remained the same; i.e. the identification of underlying phenomena, expression in a mathematical relationship, and application to a system of interest (Hill 2009).

Process systems engineering (PSE) research under these two paradigms of chemical engineering involves combining a broad knowledge of chemical products, property models, scientific experimentation, and process simulation to design efficient manufacturing processes. The usual design objective is to minimize the cost of manufacture of a product after its molecular architecture has been developed by chemists; an approach that was originally developed for commodity chemicals with simple molecular architecture produced in bulk quantities. However, as manufacturers have shifted to developing specialty chemical products to combat global pricing pressure, it is increasingly clear that this approach needs refinement. Chemical products are no longer simple mixtures of chemicals, but rather complex molecular structures with elaborate consumer interfaces that may exist at multiple length scales. This complexity means

that a sequential design approach using chemical and physical properties dependent upon both chemical engineering science and unit operations (i.e. product design, then process design) is likely to miss potential optimal configurations.

Developing systematic approaches that integrate the development chemistry with manufacturing constraints has been suggested by the Committee on Chemical Engineering Frontiers in 1988 as the third chemical engineering paradigm (Hill 2009). This new paradigm does not preclude other paradigms from emerging, or even displace the previous paradigms of unit operations and chemical engineering science, but rather it is a necessary result from breakthroughs in molecular modeling, signal processing, and increasingly powerful computational tools (Charpentier 2002). Harnessing the immense power of these new tools now allows chemical engineers to solve entirely new and important classes of problems previously unimaginable (Hill 2009). Unlike conventional process systems engineering (PSE) approaches that are focused on the design and optimization of chemical processes based on *a priori* knowledge of the products, it is now possible to investigate chemical product formulations prior to experimentation and simulation of their manufacturing processes (Hill 2009). This change in the “way of thinking” about solving chemical engineering design problems represents the beginning of a new direction in the field (Seider *et al.* 2009).

The emergence of chemical product design can also trace its roots in the United States back to the 1980's. In response to rising overseas competition from Japan in the manufacturing sector, the U.S. manufacturing base experienced a paradigm shift toward holistic design practices inspired by the Japanese concept of *Musubi* (Boznak 1993). *Musubi* literally refers to “the mutual bond of relation that makes growth possible” (Boznak 1993). In practice it shifted the focus of research from improving attributes of individual features that would be assembled in to

a product, to one that optimized the final product attributes regardless of its effect on individual features. In other words, using definitions associated more with business than engineering, chemical product design shifted from a technology-push function to a market-pull function. Technology-push occurs when a new technology is created and introduced to the market where no market demand exists. It operates under the assumption that the reason no consumer demand exists is because consumers are unaware of what is possible. Thus, simply introducing the product should create the demand. Conversely, market-pull refers to using the consumer needs and wants to drive the product development and by extension the technology development. Examples of technology-push achievements in the field of chemical engineering are new advances in polymers, drug-delivery mechanisms, and catalyst structures at the molecular- and meso-scales and crystallizers, distillation columns, and separators at the macro-scale. In addition, much work is at the micro- and nano-scales, specifically exploring the unique chemistry and physics occurring at these levels. Example of market-pull products in chemical engineering products include teeth whitening strips, post-it notes, environmentally benign refrigerants, and targeted designer drugs. In many instances it can be argued that the emergence of product design as a distinct paradigm in chemical engineering is partly due to the growth of a market-pull focus within the chemical engineering field.

Finally, the emergence of chemical product design in chemical engineering can also be attributed to the acceleration of the product development cycle (Costa *et al.* 2006). The increase in speed has put strain on the product developer's ability to maintain stability in the design process, which is key to preventing runaway costs (Boznak 1993). As a consequence, researchers have increasingly turned to computational tools to evaluate chemical products before experimentation, which gives flexibility to handling changing design constraints early in the

development process. The effectiveness of these tools can be increased by constraining the molecular architecture options evaluated. It is best achieved by first classifying chemical products by their molecular architecture, chemical and physical properties, and/or scale specific mechanisms.

2.2. Types of Chemical Products

Many options exist for classifying chemical product types. The products can be defined by their end use, by the type of raw materials used to produce them, by molecular structure, or even by the types of processing equipment needed to create them. One possibility is to define products by their physical forms. Wibowo and Ng (2002) proposes the following classification scheme:

Table 2.1: Chemical Product Types Classified by Physical Form (Wibowo and Ng 2001; Wibowo and Ng 2002).

Physical Form		Product Form	Examples
Solid	Shaped	Composites Capsules Tablets Solid Foams	Bar of soap Whale oil capsule Aspirin Tablet Styrofoam
	Bulk	Powders Granules	Powdered Detergent Pharmaceutical Excipient
Semi-Solid		Pastes Creams	Toothpaste Sunscreen
Liquid		Liquid Foams Macromolecular Solutions Microemulsions Dilute emulsions and suspensions Solutions	Shaving Foam Dishwashing Liquid Hair Conditioner Writing Ink Perfume
Gas		Aerosols	Hair Spray

Although a chemical's physical form can differentiate product types, the molecular structure between different forms can be identical. For instance, xylitol is both a pharmaceutical excipient in a solid dosage form as well as a potential biochemical platform product in liquid chemical form. The difference is the presence of strong intramolecular interactions overcoming the kinetics of the molecular motion that creates long-range order in the solid dosage form.

An alternative classification was proposed by Westerberg and Subrahmanian (2000) which listed three different classes of products based on the dominating characteristic:

1. Products which are chemicals, such as pharmaceuticals, cleaning fluids, proteins, and lubricants.
2. Products requiring chemical manufacturing techniques such as semiconductors, microfluidics, and other etching and layering technologies.
3. Products utilizing chemical properties in their functionality, such as lenses that change color in light, or a tape that sticks to surfaces, but can be removed without leaving a residue.

Costa *et al.* (2006) extends this classification scheme to include 6 categories: (1) specialty chemicals, (2) formulated products, (3) bio-based concepts, (4), transformational devices, (5) virtual chemical products, and (6) technology based consumer goods. Of particular note are the separation of functional techniques into devices that carry out a physical transformation and consumer products whose functionality is dependent on process technology. Likewise, two new categories are introduced: bio-based concepts and virtual chemical products. Bio-based concepts represent the recent growth of chemical engineering into the structural biology fields of genomics, proteomics, drug design, and membrane design. Virtual chemical products represent the software advances for chemical engineers including process design tools like ProMax®,

product design tools like ICAS®, and various open-source codes and databases like ChemSpider® and KEGG®.

Conversely, Gani (2004) classifies products based on the length-scale in which the dominant product attribute is exhibited:

1. Structured products where the micro-structural properties are related to product functions (e.g. polymer membranes, drug delivery mechanisms, etc.).
2. Chemical, agrochemical, biochemical products where the corresponding processing routes in the macro-scale play an important role (e.g. environmentally benign solvents and refrigerants, etc.).
3. Formulations where the properties are enhanced by the addition of other chemical products (e.g. solvent blends, coatings, etc.)

The use of scales to define product categories is unique, yet appropriate since the majority of products have properties that are dependent on the physical structures at these scales. Since these structures are typically interrelated, an understanding of how these scales are linked is warranted.

2.3. Chemical Product Scales

Although Gani (2004) only discusses 3 length scales in his paper, many other scales exist. In biochemistry the scales are organized in levels with increasing complexity beginning with the simplest form of biology, the gene, and ending with the most complicated, the biological production plant (Fig. 2.1). It has been shown that this multi-scale approach to biochemical design can be used to better understand and control biological tools such as enzymes and microorganisms for structured products (Charpentier 2009). Understanding the smaller scales is

becoming an intricate part of designing new products. As such, the inclusion of scales smaller than the nano-scale will also be vital to new product design. For example, the physics governing the nano- and micro-scales is often quite different from that of the macro-scale. Whereas simple chemical mixtures have been used to meet product demands in the past, now designer environmentally benign materials can be built with delivery mechanisms made possible through functionalization and self-assembly at the smaller scales. As research continues in these areas, new tools for modifying benign chemicals will be developed with multiple levels of intricacies which “offer the possibility to simultaneously engineer molecular, microscopic, and macroscopic, material characteristics” (Charpentier 2009). The ultimate goal is to efficiently identify alternative product formulations with as good or better attributes as current state of the art chemical products.

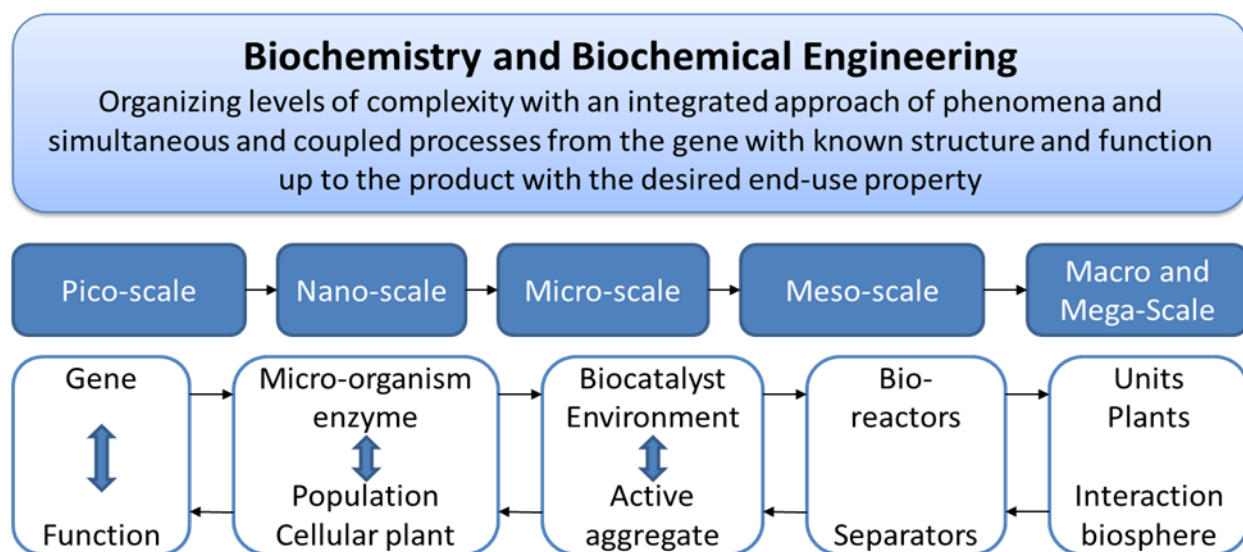


Figure 2.1. Multi-length Scale Biochemistry Approach to Product Design (Charpentier 2009).

For chemical product design and engineering, Costa *et al.* (2006) suggests that products be classified using the nano-scale, micro-scale, meso-scale, macro-scale, and mega-scales as shown in Fig. 2.2.

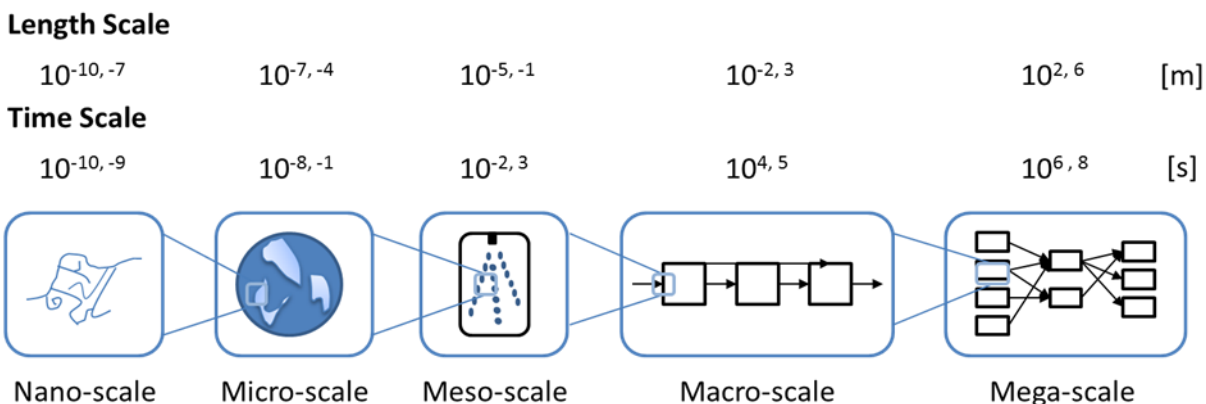


Figure 2.2. An Example Multi-length and Time Scale (Edwards 2007; Costa *et al.* 2006).

Here the classes are arranged in order of integrated length and time scales described by increasing molecular architecture complexity. In terms of the length scale, molecular surface and interaction energies are described at the nano-scale, long range order such as crystallinity at the micro- and meso-scales, processing characteristics at the macro-scale, and logistics, life cycle analysis, and consumer preference at the mega-scale. Conspicuously missing from the description offered by Costa *et al.* (2006) are length scales describing molecules. In terms of the time scales, the molecular scale is based on motion: atoms in a molecule during a reaction are measured in picoseconds and molecular vibration is measured in nanoseconds (Charpentier 2009). In this scale the importance of the molecular shape, intra- and inter-molecular forces are controlled by functional groups and their orientations. The functional groups and geometries are determined using statistical averages based on *ab initio* calculations at the quantum scale. For example, during reactions, the orientations of functional groups change considerably as molecules are pulled apart, stretched, and rearranged into new molecules. Even in a pure, thermodynamically stable state, the atoms comprising molecules are undergoing constant motion and rearrangement. The optimum, lowest energy position, is statistically the most common. It is generally accepted as the ‘normal’ shape and orientation. However, in many cases the less

probable shape is precisely what is needed for a particular product. Therefore, it is important for researchers to include these possibilities when conducting designing new products.

An attempt at a complete, multi-scale simulation figure detailing the smaller length and time scales based on common computational design approaches is given in Fig. 2.3. Of note, the nanoscale, microscale, and mesoscale have been grouped together under the molecular information demarcation because of the similarities in the computational models between them.

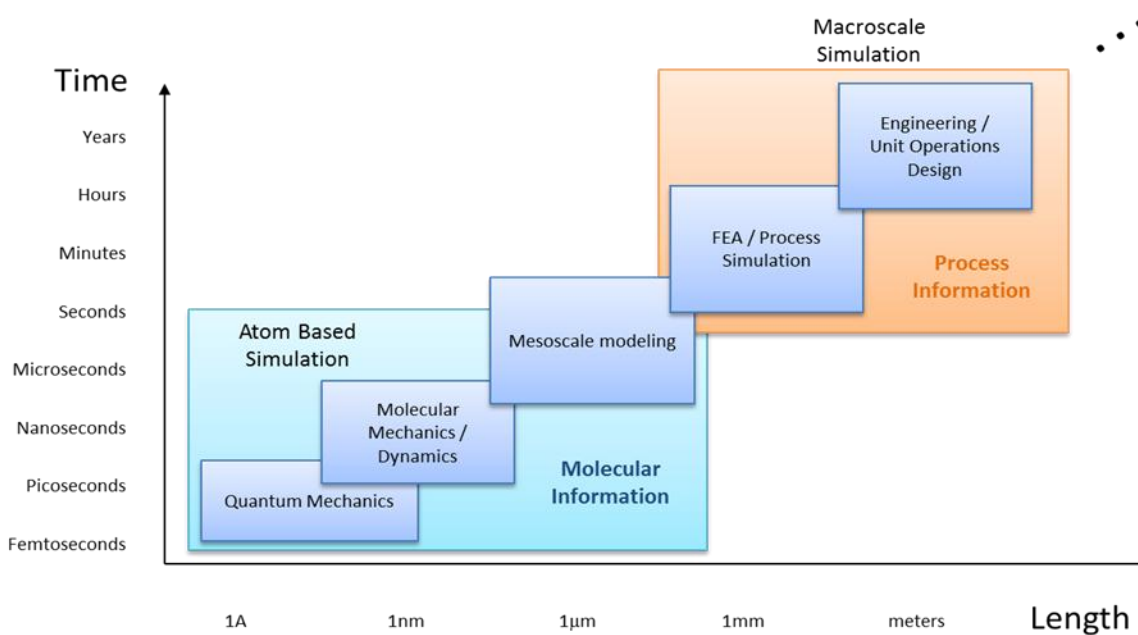


Figure 2.3: The Multiple Length and Time Scales Associated with Computational Simulations.

Alternatively, Grossmann and Westerberg (2000) proposed a product based representation based on the molecular architecture of the product, as shown in Fig. 2.4. This figure contains most of the product types typically associated with chemical engineering, such as the macro-scale unit operation and process design. However, it is increasingly clear that chemical engineering research and application will move into the smaller chemical scales. Organizing the complexity levels of the nanoscale, microscale, and mesoscale, including the

phenomenological relationships translating molecular architecture to control the functionality of products through process engineering will become vital within the discipline (Charpentier, 2005).

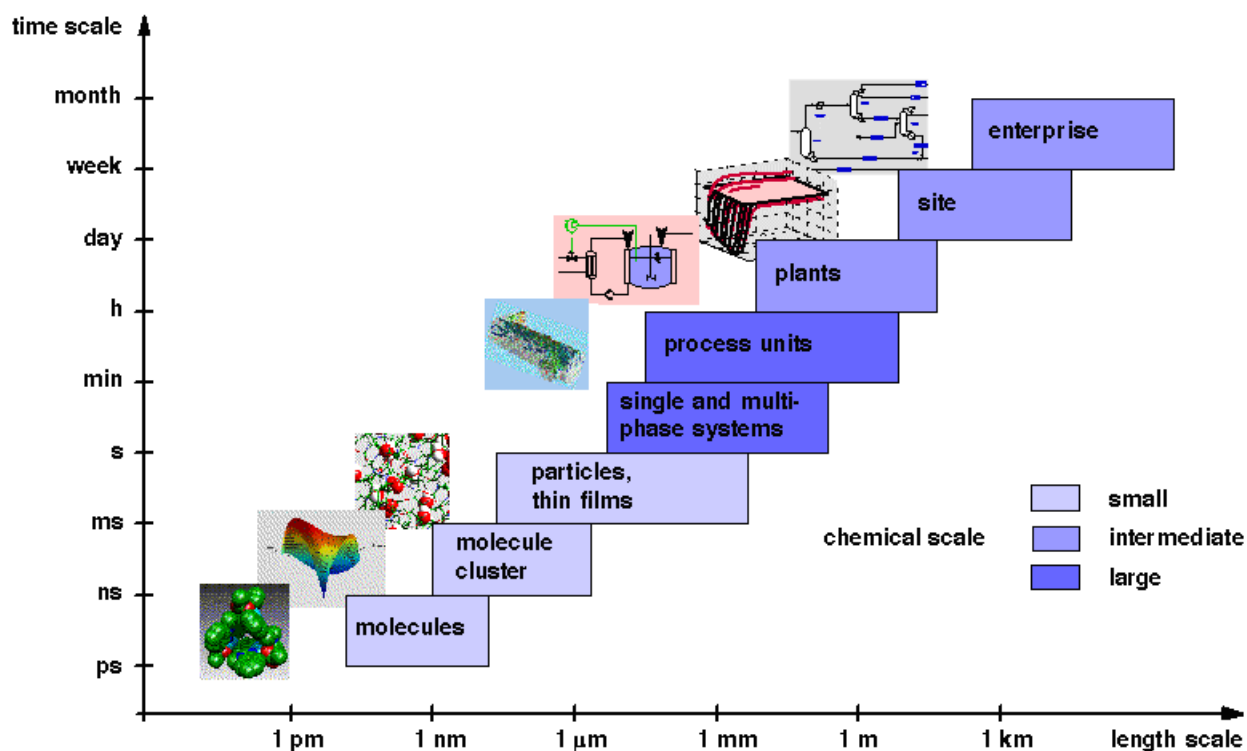


Figure 2.4: Multi-scale Overview of the Chemical Product Supply Chain Demonstrating the Potential Chemical Products Produced (Grossmann and Westerberg 2000).

2.4. Other Definitions

In the following chapters, a multiscale chemical product design approach will be presented. Some of the terms used are defined below:

- Molecular Architecture: A chemical product's molecular architecture includes information related to atoms and bonds at the molecular length scale as well as nanoscale, microscale, mesoscale, and macroscale structure information related to aspect ratio, crystallinity, morphology, and many more orientation specific arrangements of chemical structure.

- Physical-chemical Properties: Any measurable property that describes either the physical's systems state (physical property) or its change during an interaction or reaction (chemical property).
- Consumer Needs, Preferences, or Product Attributes: A set of consumer desirable product functionalities and properties that are difficult to quantify physically, such as smell, feel, or product indices, etc.
- Commodities or Commodity Chemicals: Commodities have a simple, well-established molecular architecture, are produced in bulk and have minimum variation between manufacturers.
- Formulations, Specialty Chemicals, Formulated Chemical Products: Consumer oriented specialty chemicals are made from several ingredients, produced in small quantities, have simple or complex molecular architecture, and generally have significant variation between manufacturers.
- Chemical Product Design: Chemical product design is the conversion process of a conceptual idea into a tangible, manufactured object with a defined molecular architecture.
- Design Space, Target Region: The design space is the physical-chemical property or attribute domain in which candidate molecular architectures will be investigated.

2.5. General Design Approach

In chemical engineering, product design describes the design of chemical products and may include high performance chemicals, formulated pharmaceuticals, semi-conductors, household products, beauty or personal care products, and processed foods (Hill 2009). Particularly, chemical product design involves the development of a product formulation, whose

structure is in the range of 0.1-100 μm , complete with a set of specifications for both the product and manufacturing process (Costa *et al.* 2006). It often encompasses process design, which can be thought of as a subset of the product-design activity (Seider *et al.* 2009). However, product design differs from the more traditional process design in several key areas (Westerberg and Subrahmanian 2000). First, products have short life cycles (Westerberg and Subrahmanian 2000), measured in months as opposed to years for processes. Second, the minimization of concept-to-market time can be much more important than the design of the most economical solution. Westerberg and Subrahmanian (2000) quotes that being first to market often leads to capturing and holding of a 70% of market share. My own experiences at the consumer product company Kimberly-Clark suggest this number is closer to 55-60% of the market, which is still significant. This shift in focus for designers in chemical engineering has generated a need for additional skill sets and capabilities.

The design and development methods currently used in product design consist of a mixture of talents and techniques. These techniques can be anything from loose guidelines that guide the flow of information to intricate computational logic programs that optimize specific molecular architectures for a consumer attribute. Westerberg and Subrahmanian (2000), suggest chemical product design is a mixture of tools from business, fine arts, social sciences, and traditional chemical engineering. Business tools, in particular, have a large application area in product design. Anything from deciding which raw material to buy to deciding which product to make are primarily business decisions that are part of an enormous supply chain optimization problem (Westerberg and Subrahmanian 2000). These decisions ultimately decide whether or not the product will be financially viable to manufacture. Whether or not it will sell is impacted

by the consumer side of the equation, which is dominated by the tools from the fine arts and social sciences such as understanding the appeal of color and shape.

In most instances, consumer appeal is the primary driver of performance in the marketplace. Consumer appeal is not only delivering a product with desirable attributes, which is often the sole focus in engineering, but is also the delivery mechanism itself, often called the consumer interface. For example, in 2001 Kimberly-Clark developed a new bath tissue product called Roll Wipes™. The product was a traditional bathroom roll with the addition of an applied anti-bacterial wash, resulting in a ‘wet’ tissue. Consumer research suggested there was a considerable market demand for the product. It was to be a breakthrough for the company. It was not. In fact, by most accounts, it was an abject failure. The reason was not the product itself, but because of the consumer interface. In all of the focus group testing of potential customers, the wet rolls were tested sans the product delivery mechanism that defined the product interface. After introducing the product to the marketplace, the overwhelming feedback from consumers revolved around frustration with the appearance and function of the delivery mechanism. Unfortunately, the delivery mechanism had been designed to match the constraints of the previously optimized product. Any change to the interface would mean a complete redesign of the product. Without significant changes, sales of the product reached less than one third of its projections and the product was canceled. Clearly, the importance of fine arts tools to define the consumer interface and the understanding of the interactions of the consumers is of paramount importance.

Last, but certainly not least, there is no substitute for the core competencies in chemical engineering. These tools are in essence the “chemical” part of chemical product design. Gani (2004) stresses that in chemical product design, the final product is unknown, but there are some

general ideas how it should behave and that the problem is to "find the most appropriate chemicals that will exhibit and/or cause the desired behavior." The definition of these chemical-property relationships is where chemical engineers and chemists have developed strong skill-sets, especially for fluids. However, over 50% of chemical products produced are in the solid form (Costa *et al.* 2006), so skills in engineering chemical solids are also important. Clearly, as chemical product design grows as a discipline within chemical engineering, it will need to include all of the aspects of product design, especially those associated with the solid form.

The approaches to designing chemical products can be as varied as the products themselves. However, the overall goal remains the same, to deliver the most efficient chemical product that presents a set of attributes important to the customer. As such, the approaches to designing chemical products involve the identification and transfer of pertinent information from the customer to the product's molecular architecture and vice versa. The approaches can be classified as either non-linear or linear product design.

Non-linear product design is often defined as quantum leap innovation that is unexpected and unpredictable, typically resulting from advances in technology (Boznak 1993). For example, Schmid and Smith (2002) uses Sir Alexander Fleming to illustrate that a random connection of research on the influenza virus in 1928 with 'contaminated' staphylococcus cultures exhibiting stunted growth, resulted in a miracle breakthrough, penicillin. Whereas over 50 years earlier John Tyndall had described the antibacterial properties of mold, it was Fleming who made the non-linear leap and opened "the gate for the discovery of antibiotics" (Schmid and Smith 2002). Unfortunately, as vital as these non-linear breakthroughs are to the advancement of science, they are also rare. As a result, most practical product design is often focused on only incremental or linear advancement of products.

One type of product design that uses linear transformations is presented by Westerberg and Subrahmanian (2000) and Dym and Little (2008). In this approach, product information passes from a set of market-based goals to a set of consumer attributes controlled by fundamental properties which are in turn controlled by molecular architecture. Selecting the most promising molecular architecture is a function of both its attributes and the manufacturing process used to make it. Cussler and Moggridge (2001) express this type of product design in a more objective manner by identifying four steps that capture the product design process:

1. *Needs*. What needs should the product fulfill?
2. *Ideas*. What different products could satisfy these needs?
3. *Selection*. Which ideas are the most promising?
4. *Manufacture*. How can we make the product in commercial quantities?

Cooper and Edgett (2009) further develop the product design process to include 6 business actions that emphasize interaction with the consumers:

1. *Discovery*. What new opportunities exist?
2. *Scoping*. What is the market for these opportunities?
3. *Build Business Case*. What are the potential product definitions and justifications?
4. *Development*. What is the actual product design and its business plans?
5. *Validation*. How can the product and its customer acceptance be validated?
6. *Launch*. Full commercialization of the product.

The first 3 steps listed by Cooper and Edgett are data gathering steps used in steps 4 and 5, Development and Validation, respectively (Cooper and Edgett 2009). The four of the steps listed

by Cussler and Moggridge would fall into the Build Business Case, Development, and Validation steps listed by Cooper and Edgett. Hence, the development of chemical products and their validation form a major part of the design process. In fact, it is such a major part, the tendency is to include all available information and design possibilities for the product under the false pretense that excluding some product goals could eliminate the optimum product. To mitigate overwhelming each successive step in the product design process with extraneous information, a series of gates containing criteria to measure project deliverables, are used. These criteria can be of the multi-criteria decision making (MCDM) variety, where weighted scores are assigned to the project, or they can be more explicit such as physical cost of manufacture (Parker and Moseley 2008). More details on MCDM is available in Appendix A1.

Combining the development steps with the screening stages, results in what is known as the “Stage-Gate” process for linear product design (Cooper and Edgett 2009). Shown in Fig. 2.5 is the Stage-Gate product development process as outlined by Cussler *et al.* (2010), Cooper and Edgett (2009), and Seider *et al.* (2009). The process contains a successive series of stages, where ideas and solutions are generated, followed by gates, where the solutions are screened to find those with the highest potential. Seider *et al.* (2009) lists 5 stages of which the first three are collectively referred to as chemical product design: concept, feasibility, development, manufacturing, and product introduction.

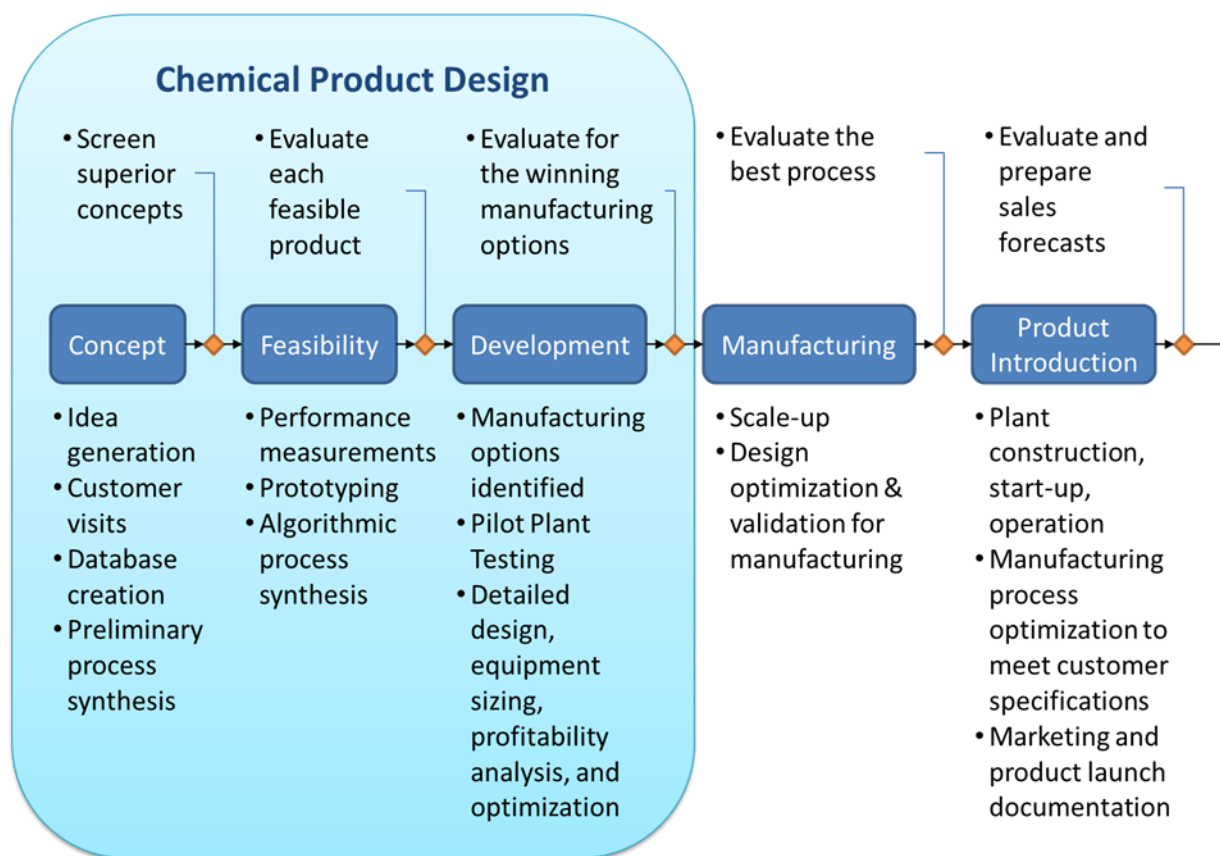


Figure 2.5: The Stage-Gate Product Development Process (Seider *et al.*, 2009).

The concept stage encompasses the initial data gathering for the product, including attributes from the consumer, potential chemical structures to deliver those attributes, and a preliminary process synthesis route for their production. The concepts are then screened in the first gate to find those solutions with the most potential. Product designs that are chosen to enter the feasibility stage are then prototyped and tested against consumer requirements from which a complete business model is built including a detailed market analysis encompassing risk assessment and health, safety, and environmental concerns (Seider *et al.*, 2009). In addition, a preliminary set of technology based production processes needed to make the product are synthesized. The overall feasibility of each process is then screened again and those with a

combination of a large potential return and small risk are passed forward to development. The development phase is where detailed design, profitability analysis, and optimization are carried out. The goal of this stage is to develop controls of the variability in the product using a set of large production runs in pilot-plants that could then be tested *en masse* in the marketplace. Those products with properties that can be controlled well are then passed forward to manufacturing and the further design of the product ceases (Seider *et al.*, 2009). The final two stages encompass manufacturing and logistical concerns with getting the product to market and will not be discussed in this dissertation.

Process systems engineering (PSE) is generally only concerned with the first three stages and gates and often tries to integrate them as much as possible in order to speed up the development time. Since significant amounts of information are generated from different sources, such as lab reports, consumer surveys, and detailed mathematical models, finding a systematic method to convert the information into a common set of criteria which can be used to provide guidance on decisions made regarding the product is paramount (Venkatasubramanian *et al.* 2006). The data, information, and knowledge rendered are usually incompatible and require several translational techniques derived from experimentation and/or mathematical programming. For consumer attributes, it is important to incorporate the specific idiosyncrasies of the different data sets and information sources that describe the molecular architecture that deliver those attributes. Historically, commodity chemical products have relied on a large amount of attribute, property, and molecular architecture information developed over years of experimentation and their design is more of a function of process development. As manufacturers have moved to produce specialty chemical products, less information is available, so chemical product designers rely on models to screen for new molecular architectures. The

data, information, and models chosen dictate the manner in which the design problem is solved. For instance, if reliable, accurate models do not exist, empirical trial and error approach based on experimentation are used (Gani 2004). One of the most common techniques is design of experiments (DOE) which uses probability to guide a series of experiments toward the desired molecular architecture. Depending on the type of molecular architecture, this technique can either be performed with conventional property parameters, attribute parameters, and/or latent property parameters based on characterization data. Further discussion on these techniques can be found in Chapter 3.

If reliable models do exist, mathematical programming or hybrid methods are preferred (Gani 2004). The choice of model is dependent on the computational efficiency and accuracy desired: more accurate, first-principle simulations are computationally inefficient while less accurate, group contribution QSPR models are computationally efficient. The key is to achieve as much accuracy as possible using as little computational resources as possible. Some attribute models are well described by physical-chemical properties and specific molecular architectures. For instance, the web strength of facial tissue is a function of the strength of intermolecular hydrogen bonding between cellulosic polymers and fibers. Developing the computer aided methods to design products with these easy to quantify attributes is often straightforward, requiring minimum effort to translate attributes into common physical-chemical properties for which QSPR relationships have already been established.

Other attributes, such as paper softness, are much more difficult to define in terms of properties and molecular architecture. Fiber strength, fiber coarseness, aloe concentration, silicone concentration, and other, undefined parameters, contribute to the attribute. For these situations, either more complex experimentation using characterization is utilized or simulation

techniques are employed. The initial focus of the PSE community for this type of design has been on the development of better simulation techniques. However, as the chemical engineering discipline moves from "data poor" to "data rich" the shift will move toward data-driven techniques (Venkatasubramanian 2009). Thus, the development of chemical product design techniques based on data and data-driven models will become increasingly important. Further discussion on computer aided approaches to chemical product design can be found in Chapter 4.

Chapter 3

Experimental Design of Chemical Products

The main objective for PSE research in chemical product design is to guide and focus experimentation (Hill 2004). Unlike in process design, the requisite properties of importance in the design of chemical products are usually consumer attributes. Consumer attributes are a set of product requirements from the consumer that have been converted into a set of consumer preferences that are empirical functions of physical-chemical properties and molecular architecture. This emphasis on end-use application implies the control of the end-use property or attribute as a primary requirement for designing chemical-based consumer products, which by extension, means controlling the molecular structure, formation, rheology, and interfacial phenomena at multiple length scales (Hill and Ng 1997; Smith and Ierapetritou 2009).

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. For situations where the set of experimental data is appreciably small, the phenomena are not well understood, or the properties of interest are uncorrelated, the attribute data is often (1) directly related to the molecular architecture or (2) related to the physical-chemical properties and then the molecular architecture. Determining an attributes relationship to the molecular architecture is performed by answering the following set of questions (Gani 2004):

1. How will the properties or attributes be obtained (calculated and/or measured)?
2. What level of detail regarding the molecular architecture will be used?
3. How will the molecular architecture be represented?
4. How will the candidate solutions be obtained?

In most cases the attributes are measured using a panel of “expert users” who have repeatedly demonstrated an ability to determine attribute differences in the desired products. Determining the molecular architecture that delivers those attributes dictates the rest of the experimental design. Due to the nature of chemistry, the smallest architectural building block that can be experimented with is a complete, thermodynamically stable molecule. Except in rare occasions, smaller building blocks, such as molecular groups, atoms, or protons/neutrons/electrons, are not stable enough to measure individually. This does not preclude the use of larger building blocks (like monomeric units of polymers, crystalline structure, particle shapes, etc.) from being designed, but does limit all experimental designs to be based on mixtures. For the first case of experimental design, mixtures of various molecular architectures are designed to directly deliver specific consumer attributes.

3.1. Mixture Design of Experiments

Design of experiments (DOE) is a form of experimental design that utilizes statistical methods to plan and execute informative experiments (Box *et al.* 1978). It is generally applied to the first stage (concept generation) and gate (screening) of the chemical product design method shown in Fig. 2.5. In some cases, process design options and constraints (feasibility and development stages) are included in DOE, relying on computational simulation and design tools to generate the process options followed by bench scale and pilot plant operations to test the options. However, since computational tools rely on physical-chemical properties to define the

system, DOE techniques are generally only used to confirm upper and lower bounds on attributes delivered by pilot-plant or larger processes.

The DOE procedure for generating and screening chemical product design concepts is as follows (Cornell 2002):

1. Postulate a model that represents the attribute or physical-chemical property response.
2. Select experimental design points to test the facets of the molecular architecture of interest.
3. Where observations can be collected, conduct experiments and fit the postulated model.
4. Test the adequacy of the model using model fitness and predicted fitness followed by calibration experiments
5. Repeat Steps 1-4 until a sufficient empirical model is developed.
6. Validate and optimize the chemical product's molecular architecture using gradient or nongradient mixture design techniques.

The most effective choice of model and location of design points is the primary focus of the experimenter. The best set of points is chosen under the following constraints: (1) the size and shape of the experimental region, (2) the number of desired experimental runs, and (3) the type of model used for constructing the map of the attribute response (Kettaneh-Wold 1991). Most often, a polynomial model is selected to represent the response surface since it can be expanded through a Taylor series to improve accuracy (Cornell 2002). A first or second degree model is usually chosen to represent the surface since it requires fewer observations. Third degree or higher ordered models are seldom utilized (Maghsoodloo and Hool, 1984). The first or second degree model may be represented as follows:

$$y = \beta_o + \sum_{i=1}^u \beta_i x_i + \varepsilon \quad (3.1)$$

$$y = \beta_o + \sum_{i=1}^u \beta_i x_i + \sum_{i \leq w}^u \sum_{w \geq i}^u \beta_{iw} x_i x_w + \varepsilon \quad (3.2)$$

Where y is the response variable and usually represents attributes (e.g. softness) or physical-chemical properties (e.g. viscosity), depending on the system of interest. x is the independent variable, usually representing the i or w molecular architecture fraction. The regressors, β_o , β_i , and β_{iw} , represent the pure component and mixture changes in y per unit change in x away from a center point. ε is the unexplained error in the response not captured by the model. The point estimate forms of the models (i.e. fitted regression models) are given below.

$$\hat{y} = \hat{\beta}_o + \sum_{i=1}^u \hat{\beta}_i x_i \quad (3.3)$$

$$\hat{y} = \hat{\beta}_o + \sum_{i=1}^u \hat{\beta}_i x_i + \sum_{i \leq w}^u \sum_{w \geq i}^u \hat{\beta}_{iw} x_i x_w \quad (3.4)$$

where $\hat{y}$ is the unbiased estimator of the response and $\hat{\beta}_o$, $\hat{\beta}_i$, and $\hat{\beta}_{iw}$, are the point estimates of the regression coefficients. To find the estimates of the regressors, the sum of the squares of the error ε (estimated at $E(\varepsilon)=0$ for uncorrelated random variables) in the model is minimized using least squares regression. Least squares regression is a common technique in the development of data driven models and is found in many forms, including classical least squares (LS), inverse least squares (ILS), multivariate linear regression (MLR), and partial least squares applied to latent surfaces (PLS). To derive the regressor solution for LS, it is convenient to switch to matrix notation. Without it, the formulas become unmanageable when the number of explanatory variables increases (Larsen 2003). Rewriting the point estimate model, shown in Eq. 3.3, in terms of matrix notation results in the following:

$$\hat{Y} = X\hat{B} \quad (3.5)$$

where $\hat{Y}$, $\hat{B}$, and X are matrices of the predicted responses, estimated regression coefficients, and independent variables, respectively, and n represents the total number of experiments.

$$\hat{Y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} \quad (3.6)$$

$$\hat{B} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_u \end{pmatrix} \quad (3.7)$$

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,u} \\ 1 & x_{2,1} & \cdots & x_{2,u} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,u} \end{pmatrix} \quad (3.8)$$

In terms of matrices, Eq. 3.5 can be rewritten as Eq. 3.9 (Larsen 2003).

$$X^T \hat{Y} = X^T X \hat{B} \quad (3.9)$$

$$\hat{B} = X^T \hat{Y} (X^T X)^{-1} \quad (3.10)$$

Solving for the best fit regressors gives Eq. 3.10 which is the central result of least squares. Although the same procedure can be extended to second order and third order models, it is seldom used because it involves fitting nonlinear models. Also, for the inverse of the identity matrix $X^T X$ to exist, X must have as many rows as columns. Since X has one row for each sample and one column for each component, then it follows that there must be at least as many

samples n as components u to be able to compute Eq. 3.10. The rules governing regression relating to samples and components are summarized by Geladi and Kowalski (1986) as the following:

- For $u > n$, there is no unique solution unless one deletes the independent variables.
- For $u = n$, there is one unique solution.
- For $u < n$, a least-squares solution is possible.
- For $u = n$ and $u > n$, the matrix inversion can cause problems.

Likewise, any linear dependence among the rows or columns of X can potentially lead to a singular $X^T X$ matrix whose inverse does not exist (Kramer 1998). This is a key observation that will be discussed in detail in Chapter 9. Other aspects of least squares regression can be found in Montgomery (2005).

Mixture design of experiments (MDOE) is an extension of DOE that explicitly uses chemical constituents as the factors in the design. By definition the constituent (or molecular architecture) fractions must sum to one and each constituent fraction must lie between zero and one:

$$\sum_{i=1}^u x_i = 1 \quad (3.11)$$

$$0 \leq x_i \leq 1 \quad (3.12)$$

Eq 3.11 imposes a colinearity effect by removing the independence of the factors. While it does not affect the utilization of the model, it does impact the interpretation of the regression coefficients. The benefit of using the constraint, however, is that it affords the ability to represent mixture data (e.g. design points, response surfaces, and target regions) in a simplex

diagram where each of the vertices represent pure chemical constituents or molecular architectures as shown in Fig. 3.1.

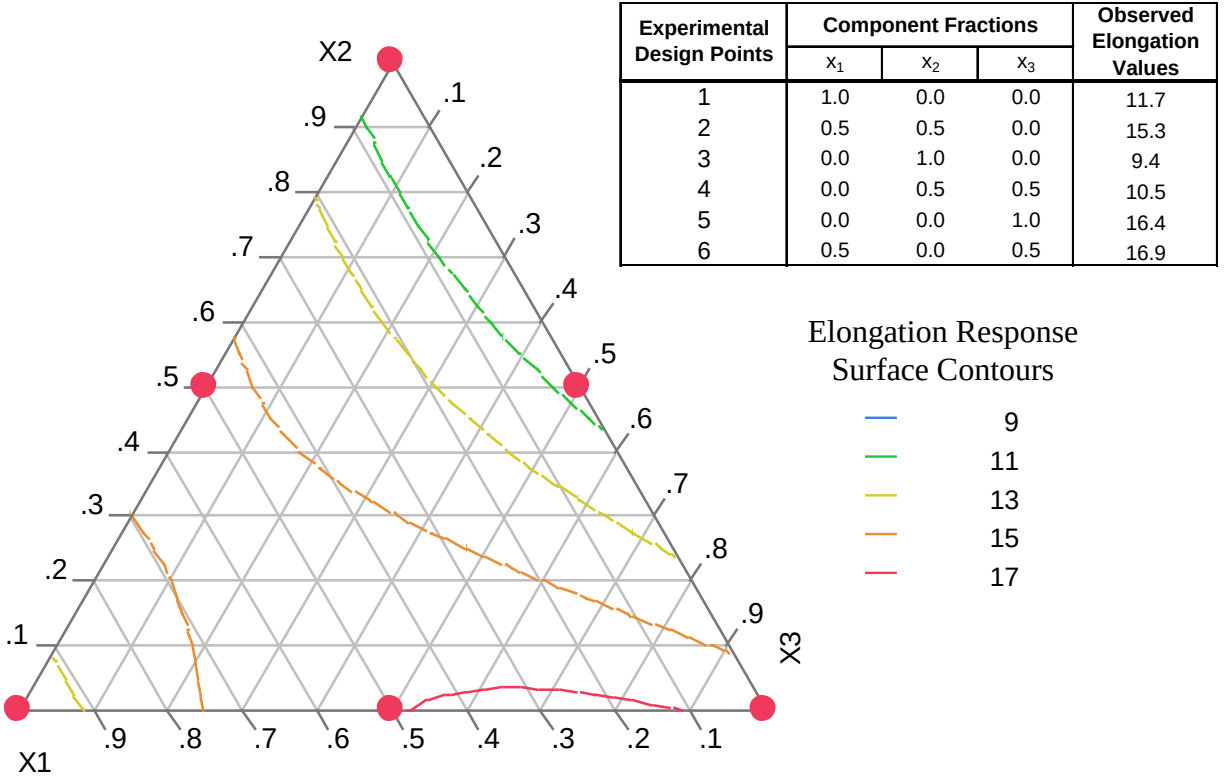


Figure 3.1: Simplex Diagram of the Response Surface of a Mixture of Polyethylene (x_1), Polystyrene (x_2), and Polypropylene (x_3) using Experimental Data from Cornell (2002).

Using Eq. 3.11, Scheffe developed the first simplex-lattice designs which many researchers considered to be the foundation of mixture design (Cornell 2002). To develop these designs, Scheffe (1958, 1963) noted that the location of the response of a mixture made up of exactly zero constituents must be identically zero meaning that the coefficient β_o is zero. Furthermore, the use of Eq. 3.11 means that the quadratic terms can be rewritten as Eq. 3.13:

$$x_i^2 = x_i \left(1 - \sum_{w=1}^u x_w \right) \quad (3.13)$$

Combining these observations, Scheffe derived the canonical models in Eq. 3.14 and 3.15. For ease of notation, the hats have been dropped from the estimates of the empirical model parameters.

$$y = \sum_{i=1}^u \beta_i^* x_i \quad (3.14)$$

$$y = \sum_{i=1}^u \beta_i^* x_i + \sum_{i < w}^u \sum_{w > i}^u \beta_{iw}^* x_i x_w \quad (3.15)$$

These first and second order canonical models are postulated to represent over 66% of all mixture designs encountered (Maghsoodloo and Hool, 1984). In situations where a higher order cubic design is desired, the same procedure may be applied to the cubic polynomial equation to make the cubic canonical model.

$$y = \sum_{i=1}^u \beta_i^* x_i + \sum_{i < w}^u \sum_{w > i}^u \beta_{iw}^* x_i x_w + \sum_{i < w}^u \sum_{w < \lambda}^u \sum_{\lambda > i, j}^u \beta_{ijk}^* x_i x_w x_\lambda \quad (3.16)$$

The effect of the canonical models is to remove the quadratic and higher terms from analysis and leave behind only the modified pure component properties and interaction effects. However, it must be noted that the modified regressors β_i^* and β_{iw}^* do not represent only pure component properties or only interaction effects. On the contrary, because of the colinearity introduced in the derivation of the canonical models, the modified regressors are amalgams of the pure component, interaction, and quadratic effects, as shown in Eq. 3.17 and Eq. 3.18:

$$\beta_i^* = \beta_0 + \beta_i + \beta_{ii} \quad (3.17)$$

$$\beta_{iw}^* = \beta_{iw} - \beta_{ii} - \beta_{ww} \quad (3.18)$$

To execute a mixture design using canonical models, Scheffe (1958, 1963) proposed using lattice design points to form a grid in the simplex diagram as shown in Fig. 3.1. Good properties of the design points consist of orthogonality, rotatability, and symmetry about an experimental center point (Lazic 2004). The design points for screening molecular architectures are located where the predicted variance of the model coefficients will be minimized, also known as minimizing *D*-optimality (Meyer and Nachtsheim 1995). For the case of the 6 design points in Fig. 3.1., the points are located at the pure component vertices and halfway along each axis. Additionally, it is common to place a point at the center of the design space, although this was not done in the example shown in Fig. 3.1. Other designs based on Bayesian *D*-optimality, *I*-optimality, and Bayesian *I*-optimality can be found in Meyer and Nachtsheim (1995). The designs are particularly useful when a constrained region is studied and no classical design is available (Eriksson *et al.* 1998).

The next step in the design is to conduct the experiments at the design points and measure the attribute and physical-chemical property responses. Since the simplex lattice design is a boundary design problem, with at least as many design points as regression terms, then the estimated regression parameters for the empirical models may be taken directly from the responses at each design point for the canonical model (Kettaneh-Wold 1992). However, since the canonical model is colinear and not orthogonal, these responses may not represent the true property or attribute estimates. The true property/attribute estimate, or orthogonal effect, is found by taking the difference in the value of the response at pure and infinitely dilute solutions while holding all other constituents constant. In addition, Cox (1971) noted that Scheffe's canonical models have the following drawbacks:

1. If two replicate experiments on the same system have the same expected responses except for a constant difference between replicates, it is obvious that fitting separate replicates, all of the regression coefficients will be different in the two replicates.
2. The absence of squared terms makes it meaningless to consider the direction and magnitude of curvature of the response to a particular component.
3. The interpretation of the regression coefficients is in terms of the responses for simple mixtures.

Likewise, Kettaneh-Wold (1991) points out that the removal of the constant term to generate the canonical model makes it impossible to center these models which often leads to an ill-conditioned $X^T X$ matrix. An ill-conditioned matrix is not symmetrical about the main diagonal, which means that the order of differentiation is important which can lead to errors during inversion. It is of the utmost importance that the constituents be independent when performing least squares regressions. To address these issues, Cox (1971) proposed a variable transformation with an arbitrary reference mixture that would allow for the use of existing polynomial models with additional constraints involving a reference mixture called the standard mixture.

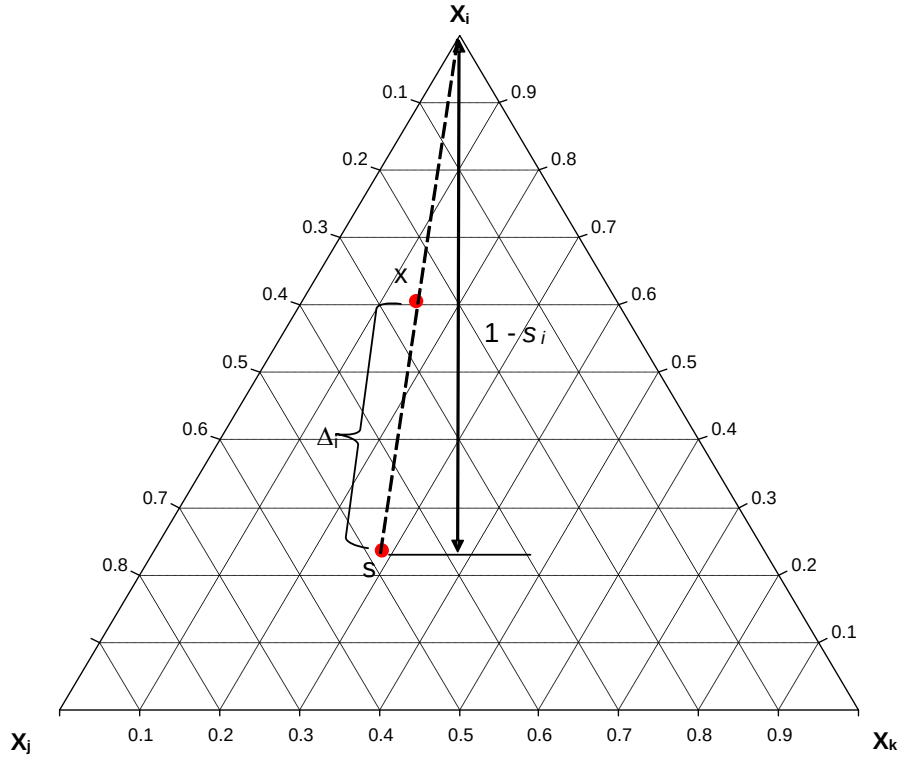


Figure 3.2: The Parameterization of the Component Space in Terms of a Standard Reference Mixture (s) such that the Incremental Change Δ_i in the Proportion of the Component i is Indicative of its Effect on the Response (Cornell, 2002).

Shown in Fig. 3.2 is a simplex diagram published by Cornell (2002) with a standard mixture s and a mixture x with a larger proportion of constituent x_i . Noting that x lies on a line from s to the x_i vertex, then the ratio of the other $u-1$ constituents are in the same relative proportions as the standard mixture.

$$x_i = s_i + \Delta_i \quad (3.19)$$

$$x_w = s_w - \frac{\Delta_w s_w}{(1 - s_w)} \quad (3.20)$$

Rewriting the first and second degree models, Eq. 3.3 and 3.4, in terms of the change in constituent i , Δ_i , gives equations Eq. 3.21 and 3.22, which provide a direct link between the regressors and the position of the design points to the reference mixture (Cornell, 2002).

$$y(x) = y(s) + \sum_i^u \left(\frac{\beta_i}{1-s_i} \right) \Delta_i \quad (3.21)$$

$$y(x) = y(s) + \sum_i^u \left(\frac{\beta_i}{1-s_i} \right) \Delta_i + \sum_i^u \left(\frac{\beta_{ii}}{(1-s_i)^2} \right) \Delta_i^2 \quad (3.22)$$

Where $y(x)$ is the expected response at design point x and $y(s)$ is the expected response at the standard reference mixture. Hence, as Smith and Beverly (1997) point out, the gradient, or change in response per unit change in x_i along the line between *Point S* and *Point X* in Fig. 3.2 is called the *effect of x_i* , provided x_i is free to range from 0 to 1. It has also been shown that the response at the standard mixture is equivalent to the estimate of β_o regressor (Box *et al.* 1978). Likewise, the transformed variables of the expressions in Eq. 3.21 and Eq. 3.22 can be represented as:

$$\sum_i^u \left(\frac{\beta_i}{1-s_i} \right) \Delta_i = \sum_i^u \beta_i x_i \quad (3.23)$$

$$\sum_i^u \left(\frac{\beta_{ii}}{(1-s_i)^2} \right) \Delta_i^2 = \sum_i^u \beta_{ii} x_i x_i \quad (3.24)$$

The end result of this analysis are polynomial expressions identical to Eq 3.3. and 3.4, but with the added benefit of the major colinearities being identified and removed for mixture design. Hence, the regressors, β_i , are estimates of the pure component physical-chemical properties or attributes. Furthermore, the estimated response surface generated by either the Scheffe canonical

models (Eq. 3.14 and 3.15) or the Cox polynomial models (Eq. 3.3 and 3.4) have been shown by Cornell (2002) to be identical. More discussion on the interpretation of these models and their parameter estimates can be found in Chapter 7.

Once the model parameters are determined using Eq. 3.10, the adequacy of the model is tested and the design is iterated until a good empirical model is developed or optimum solution is discovered. The iteration of an experimental design to find an optimum solution is among the most complex problems for a researcher (Lazic 2004). In the context of MDOE, interpretation of the models, model parameters, and the effects of each of the components on the properties or attributes of the mixture is the objective of screening designs. Beyond screening designs, the prediction of the optimum from the property or attribute models can help to focus future experimentation and improve the interpretation of the property or attribute models. The success of the design is judged by the ability to arrive at the feasibility region. Reaching the optimum is more efficient if the empirical model is adequate (Lazic 2004). The F-test is most often used as a measure of the adequacy or lack-of-fit of the model. Determining how much variance is being described by the model can also be accomplished using R^2 , a measure of fit of the model to the existing structure, and Q^2 , a measure of the predicted fit of the model when elements of the data structure are removed.

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} \quad (3.25)$$

$$Q^2 = \frac{SS_{tot} - SS_{PRE}}{SS_{tot}} \quad (3.26)$$

where SS_{tot} is the total sum of squares, SS_{res} is the residual sum of squares and SS_{PRE} is the predicted residual sum of squares. R^2 and Q^2 values of 80% and above are generally considered good. Increasing the number of model parameters is usually the preferred choice to improve the

lack-of-fit, often resulting in quadratic, special cubic, or even special-quartic models (Brandvik 1998, Brandvik and Daling 1998). For these higher order models, transformations would need to be applied in order to sufficiently linearize them in order to test their normality. A discussion on linearization techniques is presented in Chapter 6.

Once the model is deemed adequate, the region of applicability needs to be determined. Since the mixture design is performed in the local domain, then the resulting regression coefficients of the property models express factor effects exactly in that part of the domain. If the predicted optimum lies outside of the experimental region studied, then the experiments need to be recentered around the optimum. Many techniques exist to achieve this repositioning including gradient and nongradient optimization methods.

Gradient methods are based on the derivative of the response surface model and, as such, are used only when the model is deemed to be adequate. The most common gradient method is the method of steepest ascent which performs experiments at predetermined steps along the vector formed by the gradient of the response surface (Lazic 2004, Brandvik 1998, Brandvik and Daling 1998). The step size is set according to the component that has the largest effect. The method is sufficient for single property models. To handle multiple responses, the method uses weighting functions on the response surface property models to create an overall gradient function (Lazic 2004, Atkinson 1992). The weighting factors and values are at the discretion of the experimenter, which is a less than ideal situation. Also, when the number of molecular architecture options is large, then graphical representation of the method becomes difficult. It will be shown in Chapter 7 how using the property clustering algorithm and the reverse problem formulation can be used to alleviate these issues. The parameters used in the gradient expression are the regressors corresponding to each effect direction and remain unchanged as the solution is

marched toward the predicted optimum represented as the feasibility region. Once the solution leaves the feasibility region, the method is stopped and a new MDOE may be conducted. In cases where the gradient method is unable to reach the feasibility region, either the adequacy of the model needs to be improved or a more realistic set of targets needs to be specified. If the adequacy of the model cannot be improved enough to be deemed sufficient and, as such, is deemed inadequate, then nongradient optimization is performed.

Nongradient optimization searches for the optimum using a step-by-step comparison of measured property or attribute values. One of the most common types of nongradient optimizations is the simplex self-directing method (Lazic 2004). This method works by first conducting a simple mixture design around an initial guess. Next the lowest response property or attribute value is dropped and a mirror-image experiment opposite of the dropped value is conducted. The procedure continues until the algorithm repeats the same experiments, indicating that the optimum is somewhere in the bounded region. Like the gradient method of optimization, nongradient optimization struggles with marching toward a single solution for systems with multiple molecular architectures that have competing properties or attributes models (i.e. the methodology can easily become trapped in local optima). Solution methodologies to avoid such situations are discussed in Chapter 4.

Once optimization has been conducted, the design of the chemical product by the researcher is finished and process design and development begin. However, there are two major limitations in designing chemical products using MDOE that should be considered. First, while this transformation of the original Scheffe polynomials removes the primary collinearity introduced by Eq. 3.11, it leaves the secondary collinearities such as those introduced by constraints on the constituent ranges. Kettaneh-Wold (1992) suggests that the best solution may

be to refrain from interpreting the coefficients and rely on the predictions only but notes this solution is not acceptable in practice since the interpretation of regression coefficients is a necessity when the objective is to find component effects in screening situations. It is in this arena that Kettaneh-Wold (1992) suggests the use of decomposition techniques like Principal Component Regression (PCR) and Partial Least Squares on to Latent Surfaces (PLS).

In many cases because of the complexity of the interactions, especially those across multiple scales, the datasets needed to describe the molecular architecture are large and highly colinear. These datasets are typically generated using a variety of characterization techniques such as infrared (IR) or near-infrared (NIR) spectroscopy, wide angle x-ray spectroscopy (WAXS), and nuclear magnetic resonance (NMR), among others. The description of molecular architecture captured by these techniques is typically too complex to be directly described by the empirical models in a MDOE. Decomposition techniques utilize the colinear nature of large datasets to consolidate the information into a set of underlying latent variables from which mixture design can be conducted. Further discussion on experimental chemical product design using characterization and decomposition techniques can be found in Section 3.2.

Second, consumer attributes are seldom related directly to mixtures of molecular architecture in a product design because process design parameters typically dictate the use of intensive physical-chemical properties to find optimum unit-operation arrangements. Often instead, an attribute-property empirical relationship is developed to convert the attributes into physical-chemical properties, followed by chemical product design based on those physical-chemical properties. The attribute-property relationship can be captured by varying the suspected important properties in a batch of product samples and then testing them using a panel of “expert users” if they are consumer attributes, or conventional lab equipment if they are

physical-chemical properties. One of the most efficient techniques for capturing the attribute-property relationship is factorial design. Factorial design is a technique that varies quantitative (e.g. molecular weight) or qualitative design factors (e.g. softness) over an experimental design range and measures the quantitative or qualitative response of the system. The exact location of the design points depends on the base of the design. A base 2 design (2^ϕ), where the ϕ factors (i.e. a attributes or p properties) are observed at both the higher and lower bounds of the range, is most common. In base 3 design (3^ϕ), center points are added, equidistant from the higher and lower bounds. For all of the designs, the design points are located around a nexus that is as close to the expected target as possible. The number of experiments used in the design is set according to the type of model chosen to reflect the response. As before, it has been shown that polynomial models are the most efficient models for developing the attribute-property relationship (Cornell 2002).

In most cases, the number of factors influencing the consumer attributes of interest are relatively small and can be handled using full factorial design techniques. However, as the number of factors increases beyond $\phi > 5$, the number of required experiments to conduct the design becomes too large to efficiently execute. For example, a screening design involving 6 factors would require 64 experimental runs without replication or center points. For these situations, a σ fractional factorial ($2^{\phi-\sigma}$) becomes the choice in design and allows the experimenter to increase the number of factors for the same number of experiments. The trade off in using fractional factorial design is the confounding of factors with one another. With some forethought, this situation is handled by maximizing the resolution of the design. Typically, a resolution four (RIV) design is chosen that ensures that all main factors are protected from two-way interactions. Sometimes when experimental runs are constrained, resolution three (RIII)

designs may be chosen which confounds some two-way interactions with the main factors. In those cases, the experimenter may use previous knowledge to dismiss any infeasible two way interactions prior to executing the experiment.

Concurrent with developing the attribute-property relationship, a set of validation tests are used to not only confirm the existence of a suitable candidate product, but also to serve as a tool for evaluating the candidate products against a benchmark. The benchmark may be simply a comparison with a competitor's product, or they may be more explicitly defined using measurable attributes, like sales, price, etc. Most products, however, cannot be explicitly defined in terms of these attributes and need to be converted to a set of physical-chemical properties first before proceeding with design.

After the attribute-property relationship is developed and the attribute constraints are converted into physical-chemical property constraints, then the chemical product design proceeds in the same manner as it did for attributes using the steps outlined at the beginning of this section. First, the property-molecular architecture relationship is developed through a series of experiments. Then the candidate mixtures are optimized to reach a desired target. The most promising candidates proceed to the process design and development step.

3.2. Experimental Design using Multivariate Data

Often in chemical product design the molecular architecture is complex, containing important structural facets that exist over a multitude of length scales. To capture the important structural data for a single configuration requires multiple characterization experiments, each generating a large assortment of highly correlated descriptor data. Managing the complexity of this data structure requires the use of decomposition techniques such as principal component analysis (PCA), independent component analysis (ICA), and network component analysis (NCA)

to detect trends, groupings, and outliers in such systems (Wold *et al.* 1987; Gabrielsson *et al.* 2002). Decomposition techniques also afford the ability to reparameterize the system into a lower dimensional latent domain where other chemometric techniques such as principal component regression (PCR) and partial linear regression on to latent surfaces (PLS) can be used to build the attribute and/or property – molecular architecture relationships. The use of characterization property data and decomposition techniques adds an additional step to conventional MDOE blending experimental designs. An overview of the experimental design method using characterization data is as follows (Gabrielsson *et al.* 2003, Gabrielsson *et al.* 2004, Gabrielsson *et al.* 2006):

1. Identify the important molecular architecture features of the chemical product using characterization on a training set of similar chemical products.
2. Decompose the characterization data set to find the latent property domain.
3. Postulate a model that represents the attribute or physical-chemical property response as a function of latent variables.
4. Conduct a factorial design on the latent property domain to establish the attribute-latent property relationship (e.g. select the latent experimental design points to test the facets of the molecular architecture of interest, collect observations at those design points, and fit a latent variable model).
5. Test the adequacy of the model using model fitness and predicted fitness followed by calibration experiments
6. Repeat steps 1-5 until a sufficient empirical model is developed.
7. Use mixture design to identify a set of candidate mixtures based on the original training set of data that deliver the desired attributes.

8. Validate and optimize the molecular architecture of the candidate solutions using gradient or nongradient mixture design techniques in the latent variable descriptor domain.

3.2.1. Characterization Techniques

Characterization describes a class of experimental tools that gather information on not only chemical constituency or molecular structure, but also on larger structural characteristics describing the orientation and alignment of these molecules often called nano-, micro-, and meso-structure. Some examples of characterization techniques include nuclear magnetic resonance (NMR), wide angle x-ray diffraction spectroscopy (WAXS), and infrared spectroscopy (IR). The techniques are often applied to a training set of molecules defined by an experimental design used to explore the interesting facets of a set of property attributes.

Characterization, in general, provides large quantities of correlated data from which useful information on molecular architecture can be extracted. This generated data is often referred to as structural descriptors (***P***) of the molecular architecture (***X***). Fig. 3.3 has been developed as a general guide to the available options, highlighting the interconnectivity of the types of characterization, molecular architecture, and product attributes and properties available. Managing the complexity of the information requires a systematic method for determining which specific information on molecular architecture will be necessary to build appropriate models for a specific application. It is proposed that the most efficient manner for achieving this objective is to use the minimum number of characterizations necessary to generate the necessary structure information.

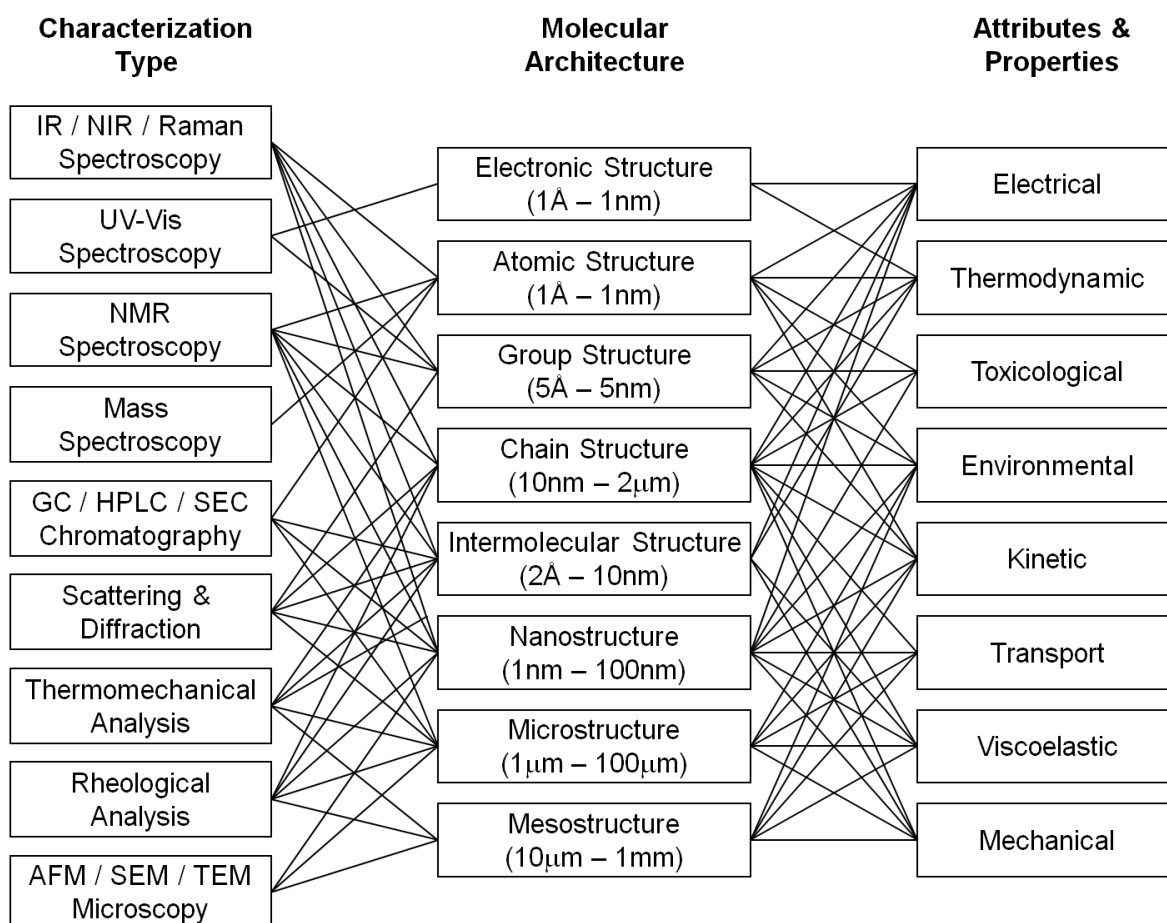


Figure 3.3: An Overview of the Interconnectivity of Characterization Techniques, Molecular Architecture, and Physical Properties and Attributes of Chemical and Material Products.

For example, Olson (1997) proposes a hierarchical structure to describe a high-performance alloy steel based on the three-link chain of processing, structure (molecular architecture), and properties as shown in Fig. 3.4 (Olson 1997; Tolle *et al.* 2009). Tolle *et al.* (2009) noted that the structures identified by Olson, namely the matrix constituency, strengthening dispersion, grain-refining dispersion, austenite dispersion, and grain-boundary chemistry could be described solely through wide angle X-ray spectroscopy (WAXS) using a model-free analysis as shown in Fig. 3.5. Other characterization and simulation options for describing the high-performance alloy steel system can be found in Olson (1997).

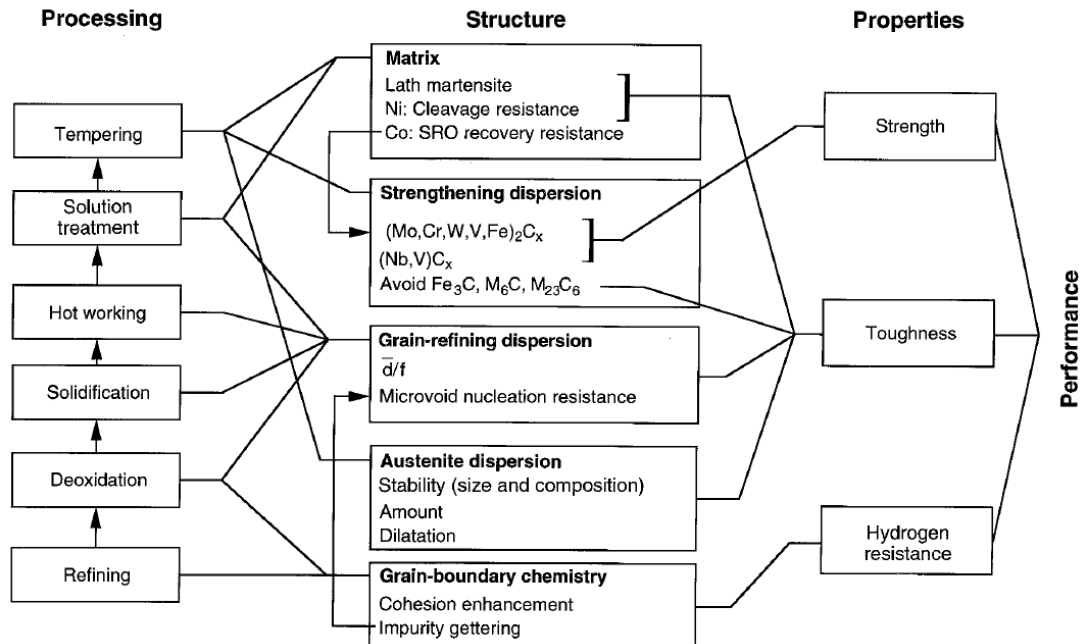


Figure 3.4: The Process-Structure-Property Relationships for a High Performance Alloy Steel (Olson 1997).

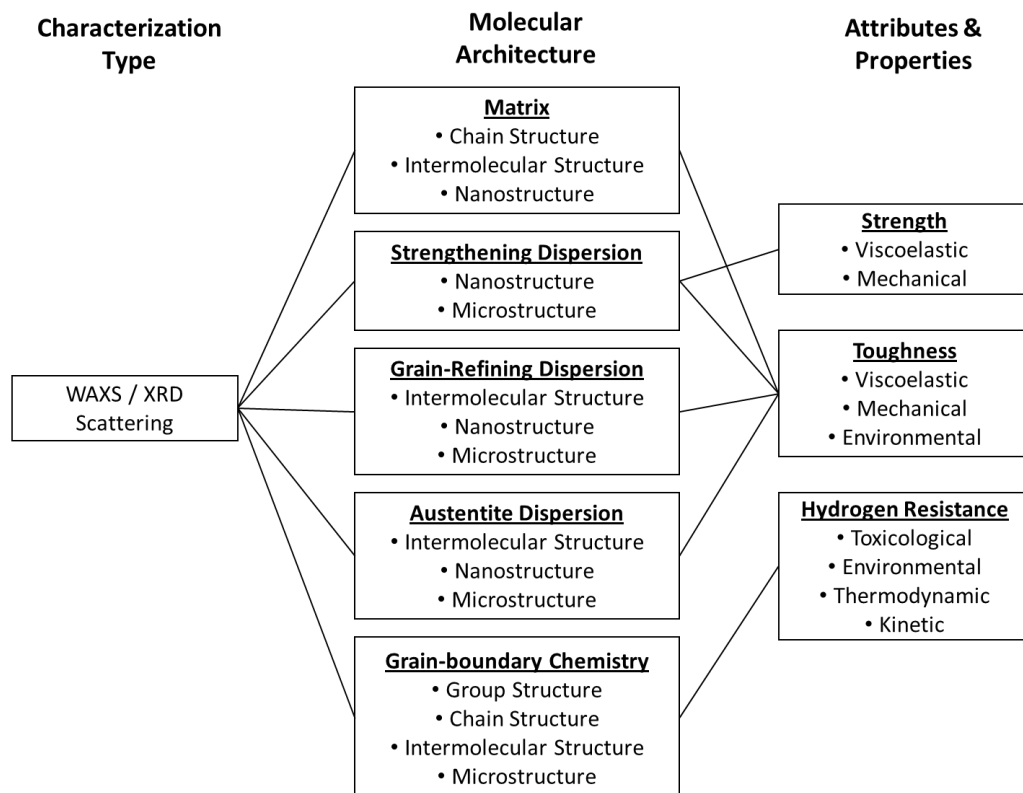


Figure 3.5: WAXS Characterization of a High Performance Alloy Steel (Olson 1997).

The main characterization used in the design of chemical products in this dissertation is spectroscopy. Spectroscopy is the detection and analysis of the radiated energy absorbed or emitted by the architecture of a chemical species (Atkins 1998). Molecular spectra are generated when the periodic motion of atomic nuclei within the molecule's architecture change when bombarded by radiation, exploiting the fact that molecules absorb specific frequencies that are characteristic of their structure. Motions can be in a straight line like symmetrical and asymmetrical stretching, and/or rotational like twisting, rocking, wagging, and scissoring (Workman and Weyer 2008). Depending on the nature of the environment surrounding the local molecular architecture, rotational, vibrational, and electronic energies are absorbed or emitted. For example, a molecule with N atoms has three degrees of freedom of motion ($3N$) with three axes of translation (x, y, z) and three axes of rotation (from inertia) (Workman and Weyer 2008). Removing the six molecular motions leaves $3N-6$ vibrational motions ($3N-5$ for linear molecules) within the molecule which can be characterized using vibrational spectroscopy.

$$N_v = 3N - 6 \text{ OR } N_v(\text{linear}) = 3N - 5 \quad (3.27)$$

The vibrational motions exist in many forms, each having a specific frequency associated with it. When the vibrations have amplitudes of 10-15% of the average distance between atoms, they are referred to as Harmonic or normal (Workman and Weyer 2008). The frequency at which harmonic vibrations absorb energy is dependent on the type of vibration, not on the amplitude. Rather, the amplitude of the absorption is determined by the absorptivity and the number of molecules encountered within the beam path. These relationships are best quantified using the Beer-Lambert Law where the absorbance P^{abs} is equated to the product of the absorptivity ε of a molecular vibration, the concentration c of the molecules in the measurement beam, and the

pathlength l_b of the beam within the measurement device. (Atkins 1998; Stephenson *et al.* 2001; Workman and Weyer 2008).

$$P^{abs} = \alpha l_b \quad (3.28)$$

In terms of transmittance P^T from 0.0 to 1.0, the absorbance can be rewritten as a function of the incident energy I_o on the sample, and the attenuated energy I after sample interaction (Workman and Weyer 2008).

$$P^{abs} = -\log P^T = -\log\left(\frac{I}{I_o}\right) \quad (3.29)$$

The magnitude of absorption depends on the concentration of the sample, which can be parameterized in terms of the number of atoms, bonds, functional groups, complete molecules, microstructure, and/or particle size and shape. Each deviation from the baseline is indicative of an aspect of the molecular architecture of the chemical product. For IR and NIR spectroscopy, the deviations from the baseline are usually a result of specific types of molecular functional groups, although in some cases, the deviations may be the result of larger scale structure (Socrates 2001, Abebe *et al.* 2008). The inherent vibrational motions of the constituent atoms within a molecular group result in unique set of frequencies at which the group absorbs incident radiation, often referred to as a spectrum. Shown in Fig. 3.6 is the infrared spectrum of five solid form excipients of mannitol.

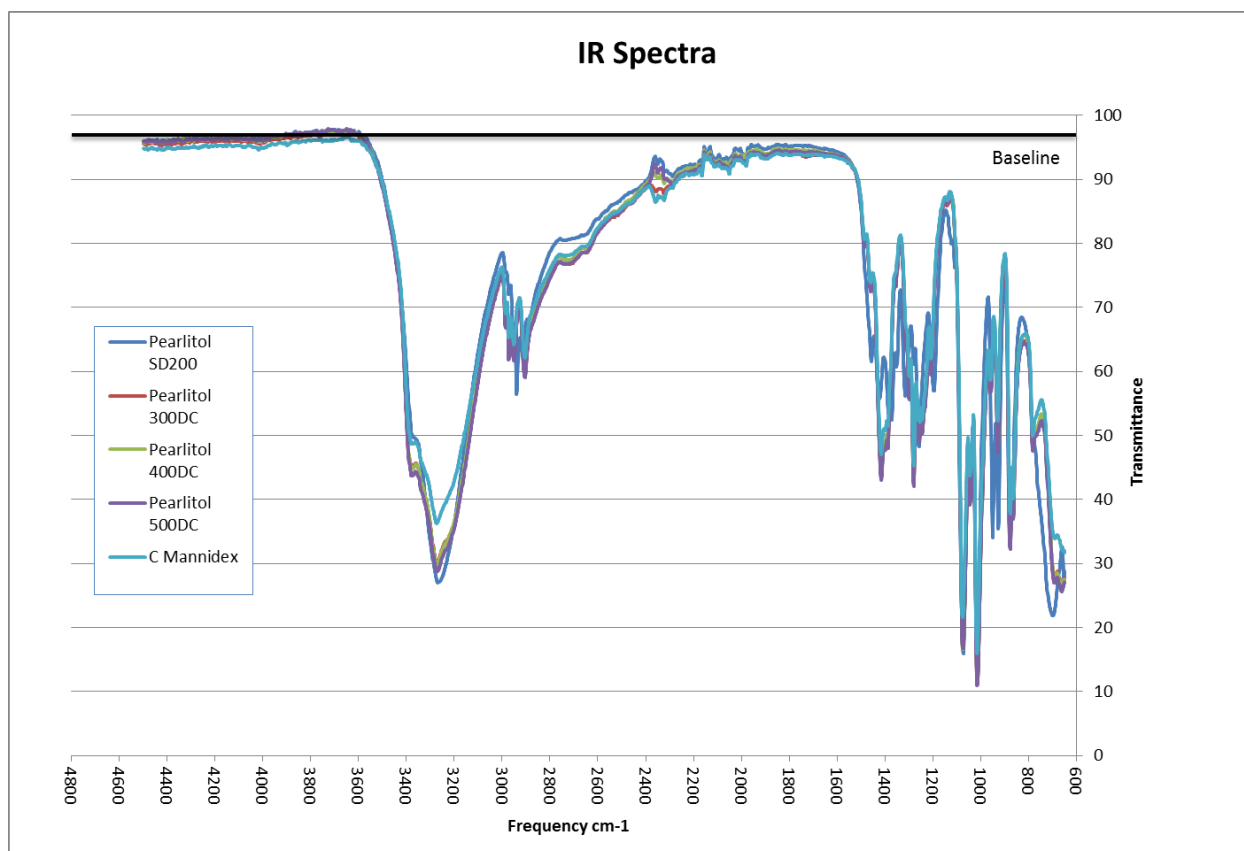


Figure 3.6: The Infrared (IR) Spectra of Different Solid Forms of Mannitol (Gabrielsson *et al.* 2003).

A molecule containing multiple groups may, in turn, contain overlaps in the spectral features, but, various “fingerprint” regions exist that can be used to differentiate between the various forms of molecular architecture (Stephenson *et al.* 2001). The most common functional groups and their corresponding frequencies have been thoroughly categorized by various authors such as Socrates (2008), Atkins (1998), Workman and Weyer (2008) amongst many others. Each functional group will often have more than one characteristic (or fundamental) absorption band associated with it; a result of different types of vibrations of which the group is capable. For example, the first order acyclic CH group has a frequency (between 2959-2849 cm^{-1}) at which the C-H bond vibrationally “stretches” and another frequency (between 1360-1320 cm^{-1}) at which the bond vibrationally “bends”. For a methylene group, 6 types of vibration include 2

types of stretching (symmetrical and asymmetrical) stretching, and all 4 types of bending (wagging, scissoring, rocking, and twisting). As the number of atoms in a molecule or functional group increases, interactions between atoms increases leading to ever increasing types of vibrational modes. For this reason, absorbance data is usually only interpreted in terms of these first six modes.

In addition to the primary fundamental absorptions, overtone and combination bands can also exist. Overtone bands result when an additional quanta of energy is absorbed by the molecule. When one quanta is absorbed, a fundamental vibration occurs. When two quanta are absorbed the first overtone is vibrationally excited, three quanta lead to a second overtone, and so on. Since the additional quanta are being absorbed, it follows that the overtone bands will approximately exist at integer multiplications of the energy level (frequency), leading most overtone bands to appear in the near-infrared (NIR) region. For instance, the first overtone band of the stretching vibration of a C-H bond is at $5917\text{-}5698\text{ cm}^{-1}$, approximately twice the fundamental vibration. Fig. 3.7 illustrates the overtone bands of the four solid forms of mannitol in the near infrared region using NIR spectroscopy.

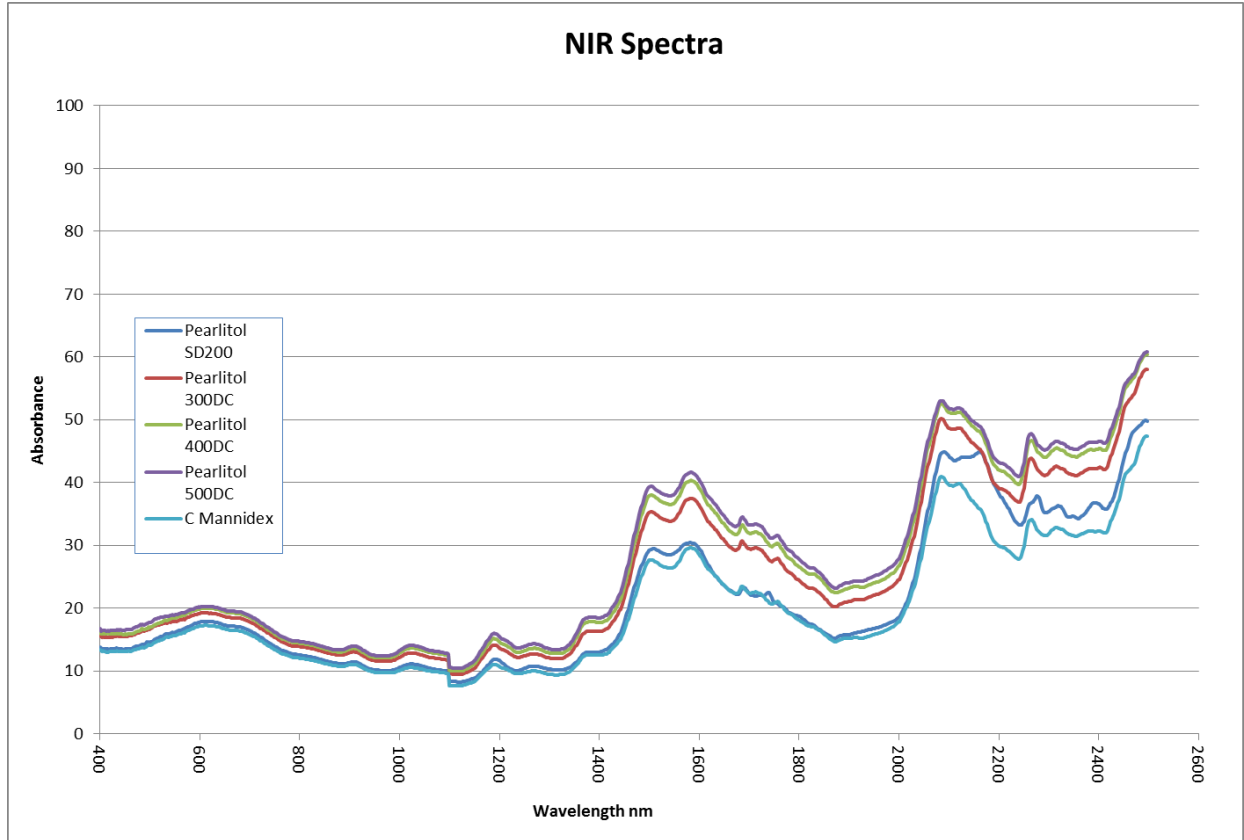


Figure 3.7: The Near-Infrared (NIR) Spectra of Different Solid Forms of Mannitol (Gabrielsson *et al.* 2003)

A more rigorous estimation of these bands can be predicted using the Schrodinger wave equation and a harmonic oscillator based on Hooke's Law according to the following expressions (Workman and Weyer 2008):

$$E_{\theta} = h\nu \left(\theta + \frac{1}{2} \right) \quad (3.30)$$

$$\nu = \frac{1}{2\pi} \sqrt{\frac{K}{\left(\frac{1}{M_1} + \frac{1}{M_2} \right)}} \quad (3.31)$$

Where K is a bond force constant that depends on the type of bond between atom, M_1 and M_2 are the masses of atoms 1 and 2, and ν is the vibration frequency, h is Planck's constant, and θ is an

integer of the vibrational quantum number (0 for ground state). The most realistic estimation of the overtone bands, however, includes a correction for anharmonic oscillation (Burns and Ciurczak 2008).

$$E_{\theta} = h\nu_{\theta}^o \quad (3.32)$$

$$\nu_{\theta}^o = \frac{\nu_1 \theta - x_a \nu_1 (\theta + \theta^2)}{1 - 2x} \quad (3.33)$$

Where ν_{θ}^o is the overtone frequency and x is the anharmonicity constant for a specific bond type as given in Groh (1998). For example, the first overtone band (2ν) for a C-H group contains a 0.019% shift due to anharmonicity ($x_a=0.00019$) resulting in a band frequency at 5917-5697 cm^{-1} .¹ An outline of the C-H stretch vibration is presented in Table 3.1 (Workman and Weyer 2008).

Table 3.1. Relative Band Intensities for C-H Stretch (Workman and Weyer 2008)

Band	Wavelength Region (cm^{-1})	Relative Intensity
Fundamental (ν)	2959-2849	1
First Overtone (2ν)	5917-5698	0.01
Second Overtone (3ν)	8873-8547	0.001
Third Overtone (4ν)	11,834-11,390	0.0001
Fourth Overtone (5ν)	14,493-12,987	0.00005

It should be noted that although group frequencies are expected to exist within narrow ranges, interference and perturbation may cause a shift in the characteristic bands due to the electronegativity of neighboring groups or atoms (Socrates 2001). These shifts can be estimated using combination bands, first order coupling, and Fermi resonance. Combination bands can

result from the confluence of fundamental bands. For instance, when a stretching vibration ν is combined with a bending vibration δ , a combination band $\nu + \delta$ results (Workman and Weyer 2008). These types of bands are generally very broad and difficult to interpret directly in terms of molecular architecture. First order coupling of vibrations occur between two or more bonds in a molecular group when the species are of similar symmetry and energy. This results in a degeneracy that splits the energy states into bands that absorb at slightly higher and lower frequencies than expected (Burns and Ciurczak 2008). Similarly, Fermi resonance is the second order coupling of vibrations and shifting of frequencies that result in an overtone band and a combination band (Workman and Weyer 2008). Group structures that contain coupled transitions are shown in Table 3.2.

Table 3.2. Group Structures Exhibiting First Order Coupling (Workman and Weyer 2008)

Absorbance Type	Group Structure	Common Examples
Cumulated Double Bond Stretching	$X=Y=Z$	$C=C=N$
Paired Stretching	$-XY_2-$	$-CH_2-$
Paired Triplet Stretching	$-XY_3$	$-CH_3$
Paired Triplet Bending	$-XY_3$	$-CH_3$
Secondary Amide Bending	$R-CO-NH-R'$	$CH_3-CO-NH-CH_3$

In Raman, additional low frequency modes are also observed. Whereas IR is adept at discerning the structure of molecules with dipole moments and NIR can capture specific functionality fingerprints, Raman can discern the structure of homonuclear systems (Burns and Ciurczak 2008). Raman works slightly differently by exciting the molecular bonds to a virtual energy state that is far above the vibrational energy levels, but is capable of describing bands

which are weak or inactive in the infrared, such as those due to the stretching of skeletal bonds such as C=C, C≡C, C≡N, C-S, S-S, N=N, and O-O (Socrates 2001). This makes Raman ideal for describing backbone structure, particularly the polarizability of symmetric structures that are inactive in IR and NIR.

Regardless of the type of vibrationally active spectroscopy (i.e. IR, NIR, Raman, etc.), two criteria form the backbone of the tools developed in this dissertation. First, it should be noted that the intensity of any spectroscopic absorbance is not only due to the strength of the bond in a functional group, but also to the number of times that bond or group occurs in the system. This observation suggests that spectroscopic information could be adapted to the group contribution method (Constantinou *et al.* 1993; Marrero and Gani 2001; Socrates 2001). A method that incorporates IR and NIR spectroscopy data into a characterization based group contribution (cGCM) approach is developed and discussed in detail in Chapter 9.

Second, to effectively use spectroscopic techniques to design chemical products it is important to capture the important features of the molecular architecture of the system in the initial training set. Although the exact number chemical constituents will vary according to the desired accuracy of design (e.g. screening vs. optimization), it is generally deemed acceptable to capture the spectroscopic characterization of approximately 30 chemical products with similar molecular architecture to the desired chemical product (Cornell 2002). Mixtures of these chemical constituents will form the initial candidate solutions for the chemical product design prior to optimization.

3.2.2. Principal Component Analysis and Regression

After spectra are generated, decomposition techniques like PCA, PCR, and PLS are used to consolidate the data, develop the underlying latent variable structure, and construct the

attribute (or property) – latent variable relationships. The objective of decomposition techniques is to describe the variation in the data using a minimum number of variables by fitting a p property characterization, structural descriptor data set of molecular architecture information to a lower, m -dimensional sub-property space or hyperplane. By using the variance-covariance structure to compress the p property data to m principal component data, it ensures that the property subspace is appropriately orthogonal and devoid of any colinearity which may exist in the property or attribute domain. The most common decomposition uses least squares as the fitting function and is known as principal component analysis (PCA) and is shown below (Gabrielsson *et al.* 2002):

$$P = T \cdot L' + E \quad (3.34)$$

Since decomposition techniques like PCA are susceptible to large differences in scales and variance, it is general practice to standardize the property data prior to analysis (Johnson and Wichern, 2007). In particular, the structure descriptor, property data matrix $\mathbf{P}$, consisting of p properties described by n molecular architectures, is mean-centered and variance scaled to unity across the p properties. This standardization helps to ensure that the multiple sources of p property descriptor data contain the same data structure and, therefore, can be decomposed into meaningful eigenvectors and eigenvalues. The underlying latent property, or score matrix $\mathbf{T}$, consists of the relative distances to the projected values of each property on to the eigenvector defined hyperplane as shown in Fig. 3.8.

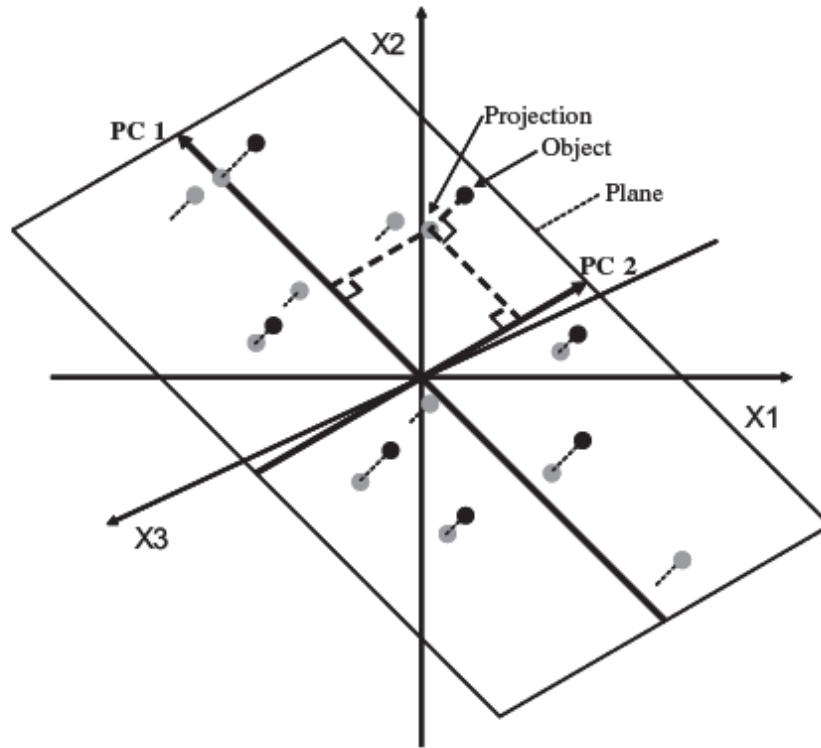


Figure 3.8: A PCA Illustration of a Least Squares Fitting of a Lower, m Dimensional Hyperplane to p Dimensional Property Data (Gabrielsson *et al.* 2002).

The eigenvalues can be considered measures of the eigenvector contrasts, or loadings $\mathbf{L}'$ of the original variables. Each loading in the loadings matrix $\mathbf{L}'$ describes how the property values are weighted together for each principal component and are represented as a hyperplane. Values of loadings vary from being highly correlated with property values, scored as -1 and 1, to being uncorrelated with values of 0. Finally, the difference between the projections and the property values are contained in the residual matrix $\mathbf{E}$. If the dimensionality of the p property domain is the same as the m latent property subdomain, then the determinant of the residual matrix $\mathbf{E}$ is 0, indicating that no error exists in the transformation. However, this condition is seldom met. Most applications of PCA result in the first 2 or 3 principal components accounting for 80% to 90% of variation p property domain, as measured by the eigenvalues (Johnson and Wichern 2007). All other information is usually found in the non-zero residuals matrix $\mathbf{E}$.

The procedure for determining the eigenvector/eigenvalue pairs in PCA, and hence the principal components, is usually performed by either singular value decomposition (SVD), spectral decomposition, or an iterative scheme like the NIPALS algorithm (Geladi and Kowalski 1986, Wold *et al.* 1987). New, orthogonal latent property components are fitted to the data structure beginning with the latent properties with the highest eigenvalues first and continuing until (1) enough variance in the data has been described (e.g. 80% or higher) or (2) the eigenvalues no longer appreciably change. A scree plot can be used to determine when the proposed decomposition is sufficient, as shown in Fig. 3.9.

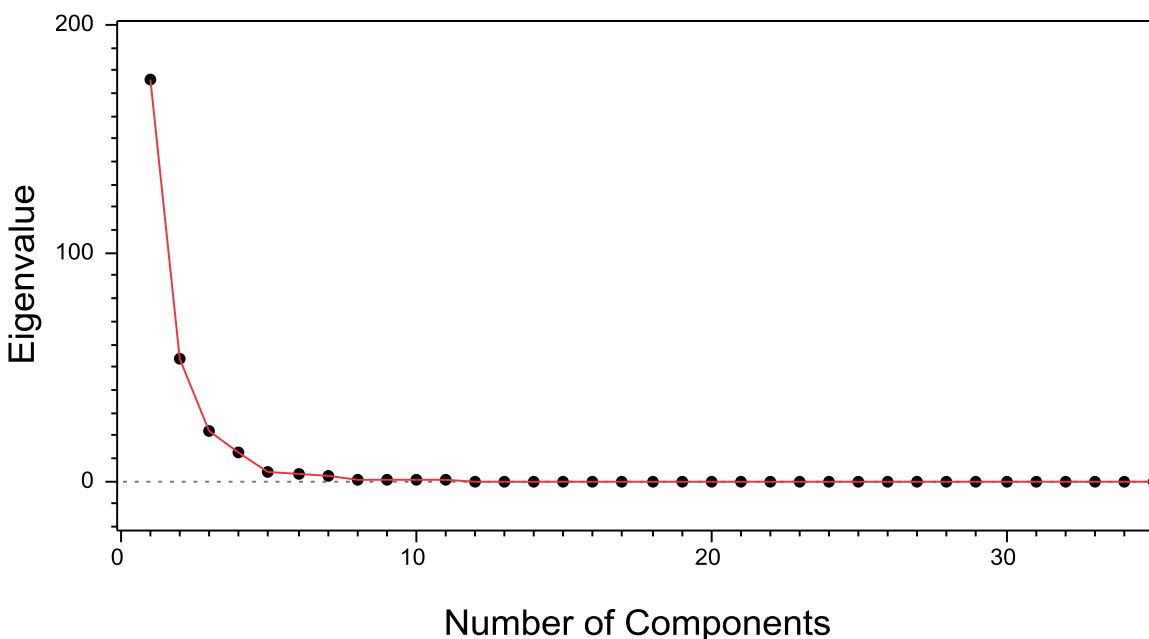


Figure 3.9: A JMP Scree Plot of the Covariance Based PCA on IR and NIR Spectroscopic Data of Filler Excipients for Direct Compressed Acetaminophen Tablets (Gabrielsson *et al.* 2003).

By definition, a scree plot is a visual depiction of the magnitude of the eigenvalues versus their number in order of decreasing magnitude. The appropriate number of latent variables are determined to be those above the bend in the plot and the remaining values are judged to be relatively small and all about the same size, meaning they are not adding significant information

to the design. These components can be removed without much loss of information (Johnson and Wichern 2007). For example, in Fig. 3.9 it was determined that 3 principal components exist above the bend and contain 91% of the variation in the characterization data (not shown), thus only $m=3$ latent properties will be used in the design. The resulting generalized PCA model is then:

$$\hat{P}_{n \times p} = T_{n \times m} L'_{m \times p} \quad (3.35)$$

Performing regression on decomposed data structures involves replacing the independent variables in the MLR expression with the underlying and orthogonal latent variables. This replacement spans the multidimensional space of $\mathbf{X}$ more efficiently when conducting experimental design which results in stronger predictive expressions. When the decomposition technique is PCA, the regression is known as PCR and has the following structure:

$$\mathbf{Y} = \mathbf{T}\mathbf{B} \quad (3.36)$$

$$\mathbf{B} = \mathbf{T}^T \mathbf{Y} (\mathbf{T}^T \mathbf{T})^{-1} \quad (3.37)$$

where $\mathbf{Y}$ is an $u \times \phi$ set of attributes or physical-chemical properties and $\mathbf{B}$ is a $m \times \phi$ set of regressors describing the attribute-latent property models. One of the greatest benefits of PCR is that because the score matrix contains no colinearity, then the $\mathbf{T}^T \mathbf{T}$ inversion will be guaranteed. At the same time, random noise is easily removed by including the top 2-3 principal score values in the model that describe 80%-90% of the data. However, to ensure that the application range of the model is suitable, it is wise to conduct any experimental design with the latent properties as the design factors. A discussion on network component analysis (NCA), which utilizes *a priori* knowledge of the molecular architecture to find the latent properties is presented in Appendix A3.

3.2.3 Partial Least Squares on to Latent Surfaces (PLS)

PLS takes PCR one step further, as it deals with the decomposition of both descriptive information (characterization property data) and response information (attribute of physical-chemical property data). Fig. 3.10 shows a PLS model being generated between descriptor data $\mathbf{P}$, which could be molecular descriptors or property descriptors, and response data $\mathbf{Y}$, which could be attribute or property information. For example, suppose a PCA is performed on both the $\mathbf{P}$ and $\mathbf{Y}$ data, represented as blue arrows in Fig. 3.10. The descriptor scores u and plotted against the response scores t , and the resulting least squares fit between the data sources is the PLS model. This approach ensures the best possible correlation between the two data sets, but not necessarily the best description of the $\mathbf{P}$ and $\mathbf{Y}$ data individually.

Geladi and Kowalski (1986) favor an iterative algorithm called NIPALS to handle the development of these relationships. The technique uses relational forms to build outer relations (PCA) of $\mathbf{P}$ and $\mathbf{Y}$, inner relations between $\mathbf{U}$ and $\mathbf{T}$ and a least squares regression for the mixture of the two. It is important to recognize that to maintain orthogonality, loading weights will also be needed. The generalized approach is as follows:

$$\hat{\mathbf{Y}}_{n \times \phi} = \mathbf{U}_{n \times l} \mathbf{W}_{l \times \phi} \quad (3.38)$$

where $\hat{\mathbf{Y}}$ is again the point estimate of response (i.e. attributes or physical-chemical properties), $\mathbf{U}$ is a $n \times l$ matrix of the latent variable scores and $\mathbf{W}$ is a $l \times \phi$ matrix of the latent variable loadings. As with the property descriptor latent properties, the latent variable representations of the attributes will contain a significantly lower dimensionality. Regressing the lower dimensional $\mathbf{U}_{n \times l}$ attribute scores against the $\mathbf{T}_{n \times m}$ property descriptor scores of Eq. 3.35, results in Eq. 3.39.

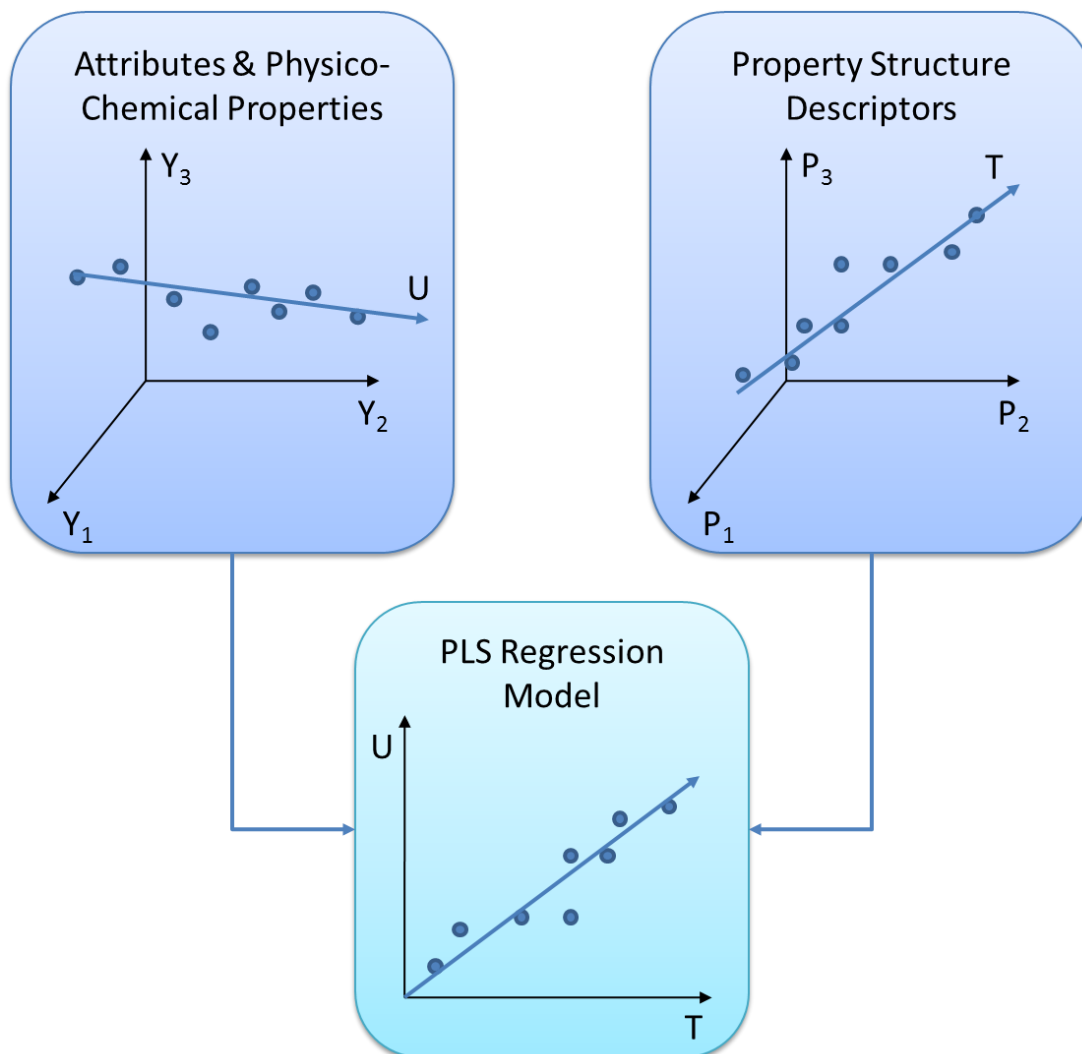


Figure 3.10: A PLS Regression Performed on the $\mathbf{P}$ Descriptor and $\mathbf{Y}$ Response Variables (Gabrielsson *et al.* 2002).

$$U_{n \times \phi} = T_{n \times m} B_{m \times \phi} \quad (3.39)$$

where $\mathbf{B}$ is a set of $m \times \phi$ regressors describing the latent attribute-latent property descriptor relationship. Iterative algorithms like NIPALS are utilized to determine the latent attribute variables, latent property descriptor variables, and regressors, A detailed description of how the

method calculates the latent variables can be found in Geladi and Kowalski (1986), Wold *et al.* (1987), and Muteki *et al.* (2006, 2007).

3.2.4 Mixture Design and Optimization with Latent Variables

Finding the optimum mixture or set of candidate mixtures that deliver a set of consumer attributes or physical-chemical properties is the customary objective of the chemical product designer. Muteki *et al.* (2007) outlines four distinct problem formulations when using multivariate data: (1) the design of blend fractions where the molecular architecture training set is predefined, (2) the design and selection of undefined molecular architectures using a database, (3) the design of blend fractions with a predefined molecular architecture under process constraints, and (4) the design and selection of molecular architectures and process parameters. The focus of this dissertation is on improving the first two cases, leaving the latter two cases for future work.

For Case 1, traditional mixture models like Scheffe canonical and Cox polynomial models are usually employed to investigate the relationship between the molecular architecture blend fraction $\mathbf{X}$ and the final product attributes and/or properties $\mathbf{Y}$. Kettaneh-Wold (1992) and Eriksson *et al.* (1998) propose using PLS regression to fit both Scheffe linear and Cox polynomial models for mixture data and show how it can effectively deal with colinearities. The proposed mixture models are typically linear in nature, but recent work by Muteki and MacGregor (2007) has shown that nonlinear relationships can improve accuracy in some cases. However, although this approach can deal with mixture colinearity and provide good estimates of the model parameters, it is limited to describing attribute-molecular architecture directly which limits the chemical product design to only be a mixture of the chemical constituents used in the training set.

Case 2 is a recent exploration within the process systems engineering field and addresses the shortcomings found in Case 1. Gabrielsson *et al.* (2003), Muteki *et al.* (2006, 2007), Garcia-Munoz *et al.* (2010), and Solvason *et al.* (2008, 2009, 2010), all propose using a combination of attribute-latent property and latent property-molecular architecture relationships to design chemical products in the latent domain. The attribute-latent property relationship can be derived from an attribute property descriptor relationship using either PCA or PLS.

$$Y = f(P) + E_y \quad (3.40)$$

Any nonlinearity in the attribute or physical-chemical property response should be handled in this expression (Muteki and MacGregor 2007). The latent property-molecular architecture relationship is derived from the property descriptor-molecular architecture relationship in the same manner:

$$P = f(X) + E_p \quad (3.41)$$

The most common relationship is ideal linear mixing:

$$\hat{P}_M = XP \quad (3.42)$$

which can be reduced to a mixture of the latent score variables using decomposition.

$$\hat{T}_M = XT \quad (3.43)$$

It should be noted that the loadings (L) are treated as transformation functions between the latent domain (T) and the property descriptor domain (P). Other types of nonlinear mixing have been explored by Muteki and MacGregor (2007), but are beyond the scope of this dissertation.

To solve the design problem, Gabrielsson *et al.* (2003) suggests using a training set of experiments to decompose the design space into latent variables first, followed by conventional mixture design to find the training set candidate with a latent variable structure that matches the

optimum predicted structure. Experimental gradient and non-gradient optimization techniques are then employed to explore the design space. Recently, Muteki *et al.*(2006, 2007) and Garcia-Munoz *et al.* (2010) proposed a similar method that searches a much larger database of potential chemical products to find latent variable matches of the original training set latent variables that deliver the desired attributes or physical-chemical properties. This approach makes the assumption that the attributes and/or physical-chemical properties of the chemical products in the database correlate well with the training sets properties. The same assumption also applies to the molecular architecture. Both of these conditions are naturally met in industrial applications because of the importance of blending fractions, but require verification in a laboratory setting (Muteki *et al.* 2006). A general discussion on the mathematical formulation of this optimization problem is presented in Chapter 4, and a specific example is provided in Appendix A7.

There are several advantages to using the two stage approach in Case 2 over the single stage approach of Case 1. First, it provides a direct interpretation of the effect of the property descriptors ($\mathbf{P}$) on final product attributes and physical-chemical properties ($\mathbf{Y}$), that is, it provides information on what property structure descriptors affect what attributes, something that traditional mixture models using only the $\mathbf{X}$ and $\mathbf{Y}$ matrices struggle doing (Muteki and MacGregor 2007). Second, Gabrielsson *et al.* (2003) and Muteki *et al.* (2006, 2007) have shown that the two stage approach provides better estimates of the blend attributes and/or physical-chemical properties ($\mathbf{Y}$) than the traditional mixture models because of the use of the additional structure descriptor information ($\mathbf{P}$). Third, Garcia-Munoz *et al.* (2010) and Muteki *et al.*(2006, 2007) have shown that new chemical products that have never been used in the past can be found by matching the latent structure of the designed mixture ($\hat{\mathbf{T}}_{\mathbf{M}}$) using a mixture design and optimization framework (Muteki and MacGregor 2007).

However, while a powerful addition to the tool set available to the chemical product designer, the two stage method presented in Case 2 suffers from the same limitations all database searches do; namely, the design is only as good as the chemical products available in the database. Logically, if the database is ill-formed or incomplete, the “optimized” set of chemical product candidates will only be local estimates of the true, global optimum candidate. This observation is discussed in more detail in Chapter 4.

3.3. Summary

In conclusion, experimental design approaches have traditionally been the choice of chemical product designers when understanding of the molecular architecture, consumer attributes, or physical-chemical properties are poor. They offer real, tangible results and instant validation of the process needed to manufacture the chemical product. While many advances have been made to improve the efficiency of experiment based chemical product designs, particularly for optimization, the designed candidates will always be subject to the ability of the training set to capture the appropriate relationships between the molecular architecture, the attribute, and physical-chemical properties. Hence candidate solutions generated with experimental based design and optimization techniques will be subject to some uncertainty. In most cases this is an acceptable solution, however for some classes of problems computer based design and optimization methods are preferred. A discussion on computer based methods can be found in Chapter 4.

Chapter 4

Computer Aided Design of Chemical Products

Chemical product design and development involves the conversion of a conceptual idea into a tangible, manufactured object with a defined molecular architecture. Design is sometimes also referred to as formulation, a term typically associated with the pharmaceutical industry that describes the method of combining different spatial arrangements of chemical species to make a medicinal product. It encompasses the generation of a concept, followed by the gathering of many sources of information, and the conversion of that information into a product with detailed performance criteria including a rigorous process design (Venkatasubramanian 2008, Zhao *et al.* 2006). Due to the immense amount of information needed to make a targeted, stable specialty chemical product, and noting that this information is often inter-related, formulation is often approached using mathematical logic programming. Many types of product formulation programs exist, each incorporating specific idiosyncrasies of different data sets, information sources, and length-scales. To solve these types of problems, a mixture of talents and techniques is needed to acquire information and distil it into useful programming constructs. Historically, performing this task has been the study of computer engineers and scientists, but increasingly, as chemical engineering moves from a 'data poor' to a 'data rich' discipline, integrating information flow methods into conventional chemical engineering design procedures will become increasingly important (Venkatasubramanian 2009).

4.1. Mathematical Formulation of Chemical Product Design

A mathematical view of the chemical product and process design problem is presented by Gani (2004). :

$$F_{OBJ} = \max\{C^T d + f(x)\} \quad (4.1)$$

$$h_1(x) = 0 \quad (4.2)$$

$$h_2(x) = 0 \quad (4.3)$$

$$h_3(x, d) = 0 \quad (4.4)$$

$$l_1 \leq g_1(x) \leq u_1 \quad (4.5)$$

$$l_2 \leq g_2(x, d) \leq u_2 \quad (4.6)$$

$$l_3 \leq Bd + Cx \leq u_3 \quad (4.7)$$

In this formulation, the chemical product design objectives are defined by an objective function (Eq. 4.1) and set of constraints (Eq. 4.2 – 4.7). For convenience, the molecular architecture information ($\mathbf{X}$) has been broken into $\mathbf{x}$, a vector set of continuous variables representing mixture compositions of molecular architectures, and $\mathbf{d}$, a vector set of binary integer variables identifying the presence of molecular descriptors such as atoms, molecular groups, and other types of product architectures. This differentiation in molecular architecture improves the efficiency of the algorithms used to solve the formulation and is discussed in more detail in the next section. In a similar fashion, the relationships between the product's attributes and physical-chemical properties ($\mathbf{Y}$) are differentiated using the following scheme: $h_1(\mathbf{x})$ are shown as a set of attribute or property equality constraints related to process design parameters (e.g. pressure, reflux ratio), and explicitly described by the chemical composition, $h_2(\mathbf{x})$ are the set of

attribute or property equality constraints explicitly described in terms of process models like mass and energy balances, $h_3(\mathbf{x}, \mathbf{d})$ are a set of attribute equality constraints related to molecular structure, composition, and other information regarding product architecture, $g_1(\mathbf{x})$ are a set of compositionally dependent inequality bounds on the process design specifications, and $g_2(\mathbf{x}, \mathbf{d})$ are a set of composition and molecular architecture inequality bounds on the product. Linear inequality constraints are described explicitly by Eq. 4.7. It should be noted that the binary descriptor variables ($\mathbf{d}$) may differ depending on where they are used in the mathematical formulation. When used in Eq. 4.1, descriptors represent the presence or absence of unit operations while in Eq. 4.4, the descriptors represent molecular architecture features. Finally, the term $f(\mathbf{x})$ represents a vector of linear or non-linear attribute-chemical composition objective functions of the system. A detailed list of the variations of the formulation are listed below (Gani 2004):

1. *Satisfy only Eq. 4.6.* This type of formulation represents a database search for validated molecular structures that satisfy the property constraints.
2. *Satisfy only Eq. 4.4.* This type is referred to as enumeration, where all candidate molecules are generated.
3. *Satisfy Eq. 4.4 and 4.6.* There are two methods for solving this type of formulation. The first is to independently apply constraint 4.4 and then 4.6 in series in a process known as “generate and test.” The second type of formulation is to apply constraints 4.4 and 4.6 in parallel. In this type, only molecular architectures that meet constraint 4.6 are generated. This is known as a “reverse problem formulation” and will be discussed in more detail in a Chapter 6.

4. *Satisfy Eq. 4.1, 4.2, and 4.6.* This type of formulation identifies the optimum molecular architecture that satisfies the product constraints.
5. *Satisfy Eq. 4.2 – 4.7.* This type of formulation finds all molecular architectures that satisfy not only product attribute constraints, but also process model constraints simultaneously.
6. *Satisfy all equations.* This formulation finds the optimum molecular architecture that satisfies both product and process constraints simultaneously.

When the formulation only contains linear functions, the problem can be performed by commonly available programming techniques in familiar software such as Visual Basic for Applications (VBA), MatLab, and/or LINGO. However, when the formulation contains nonlinear functions, the solution to the problem is more complicated and may require specialized solvers such as those found in GAMS. Furthermore, the methods listed by Gani (2004) for solving the chemical product design problem can be distilled into 3 basic types: searching a database (Variation 1), generating a candidate (Variation 3 or 5), and optimization (Variation 4 or 6). The first is related to searching for candidate molecules through a multitude of databases.

4.1.1. Searching a Database

By definition, a database is a large collection of data that is organized so that its contents can be easily searched and retrieved, usually by a computer. Ugi *et al.* (1993) notes that database searches are one of the four categories of chemical product design, according to definitions used by chemo-informaticists. The other categories, computer-assisted structure elucidation (CASE), computer-assisted synthesis-design (CASD), and 3D structure builders, focus on generating molecular architectures to meet consumer attribute targets, optimizing the molecular architecture,

and developing the graphical user interface (GUI), respectively. CASE and CASD are discussed in detail in Section 4.1.2 and 4.1.3. The last category, 3D structure builders, will not be discussed in this dissertation.

Database searches rely heavily on the creation of a comprehensive database containing information on molecular architecture, physical-chemical properties, and attributes. Historically, a chemical product's molecular architecture has been defined using characterization techniques like infrared (IR) and mass spectroscopy and these were the first sets of data to be combined in a database (Zemany 1950). To search the data, Ray and Kirsch (1957) developed the first structure search algorithm based on atom-to-atom connectivity indices. Later, Gluck and Morgan (1965) developed the first canonical form of the connection table using both atom-by-atom and bond-by-bond descriptions. The American Society of Testing and Materials (ASTM), the Chemical Abstract Service (CAS), the National Institute of Health (NIH), and many more organizations have all developed versions of spectral databases and retrieval systems. Obviously, searching all of these databases with user specific codes would become quite cumbersome which is why expert systems utilizing computer code to call up specific functions have been developed. The use of domain-limiting steps within the expert system further reduces the computational complexity by focusing in on molecules that meet certain physical criteria. An example of a database search expert system is presented below.

Molecular transport through the cell membranes is an important biological function. In most cases the hydrophobicity of a molecular compound determines whether or not it will permeate a cellular membrane (Wang *et al.* 1997). Hydrophobicity and other biological activities have been related to the octanol-water partition coefficient ($\log K_{ow}$) via quantitative structure-activity relationship (QSAR) studies. A database of physical property data, including

the partition coefficient, can be searched using the software BioPath.Explore (Ertl 2010). It contains the physical property data important in biochemical pathways for over 1,175 molecular structures and 1,545 transformation reactions. Suppose it is important that no molecules are introduced to our system with partition coefficients larger than -1.95. The database is searched and only 4 potential molecules are found to meet this criterion, as shown in the following table.

Table 4.1. Chemical Structures Provided by BioPath.Explore that Meet $\log K_{ow}$ Constraints (Ertl 2010).

Record	Formula	Name	LogK _{ow}
3	C ₂₆ H ₄₀ N ₇ O ₁₈ P ₃ S	(2S, 3S)-3-Hydroxy-2-methylbutanoyl-CoA	-6.921
9	C ₂₅ H ₃₆ N ₇ O ₁₉ P ₃ S	(R)-2-Methyl-3-oxopropanoyl-CoA	-5.032
16	C ₆ H ₁₁ O ₁₀ P ₂	(R)-Mevalonate-5-diphosphate	-1.977
48	C ₃ H ₈ O ₂	1,2-Propanediol	-2.817

The first two results are coenzymes and the third contains diphosphate. If additional constraints are added to the solution such that only carbon, oxygen, and hydrogen are utilized in the solution, then only 1,2-propanediol becomes acceptable.

The potential drawback to this type of database search is obvious: the solution is never complete (Ugi *et al.* 1993; Ugi *et al.* 1994); as new chemicals and reactions are added to the database, the old solutions can no longer be considered optimal. Vladutz (1963) and Ugi (1994) note that the chemical horizon of these programs is limited to those parts of chemistry that are stored in databases. This is evidenced by the fact that BioPath.Explore currently stores only 20% of the reactions identified in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Although this problem is somewhat mitigated by open-source database communities working to

network as much data as possible, like POPE (Akkisetty *et al.* 2009), the method would benefit from a more systematic approach to searching.

Likewise, databases are also subject to the quality of the information that is loaded into them. For instance, when the octanol-water coefficient, $\text{Log}K_{ow}$, for the 1,2-propanediol is cross-referenced with other databases, such as ChemSpider (Williams 2007), a different property value of -0.92 is obtained. Thus, using database searches by themselves have a large capacity to miss suitable molecular architectures that deliver physico-chemical properties and attributes desired by the consumer. An alternative approach is to generate the database from a smaller set of validated property structure descriptors.

4.1.2. Generating Candidate Solutions

The use of computers to solve the chemical enumeration problem, often referred to as synthesis or design, is a long standing tradition in chemistry. Early works by Corey and Wipke (1969), first showed that products and their synthesis pathways could be developed using logic schemes for the computer combined with “information controllers”, or heuristics, in the hands of the researcher. The logic portion involves an automated procedure of enumerating the chemical structure using chemical knowledge (e.g. structural descriptors). The absolute minimum input into a logic based enumerator is the knowledge of the number of atoms, with no specification regarding the atom type, etc. For instance, there are 24 atoms in d-glucose. Using only the knowledge that these 24 atoms may be arranged in any order, then $24! = 6.024 \times 10^{23}$ molecular structures would be feasible (Ugi *et al.* 1993). A significant reduction in the number of possible atomic arrangements can be obtained by applying chemical knowledge to the problem, beginning with rules about atomic nuclear structure (e.g. d-glucose only contains atoms of carbon, oxygen, and hydrogen). It significantly reduces the problem such that $N_C!N_O!N_H! = 6!6!12! = 2.483 \times$

10^{14} possible candidates remain. A further reduction in the possible candidates can be obtained by including information on the valence shell electrons involved in bonding. According to molecular graph theory, the total sum of valencies of all vertices in a graph equals twice the number of edges plus rings (Trinajstić 1992). Eljack *et al.* (2007) defined this observation in terms of the number of unused bonds in a generated molecule called the Free Bond Number (*FBN*).

$$FBN = \sum_{\tau=1}^{\alpha} n_{\tau} FBN_{\tau} - 2 \cdot \left(\sum_{\tau=1}^{\alpha} n_{\tau} - 1 \right) - 2 \cdot N_{RG} \quad (4.9)$$

Here, τ is the atom type, α is the total number of atom types, n_{τ} is the number of atoms of type τ , and N_{RG} is the number of rings. 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 configurations of the atoms in the molecule. Applying this criterion and constraining the problem to default molecular size parameters listed in ProCAMD in ICAS 13 (Gani *et al.* 2010), and constraining the problem to acyclic compounds, reduces the potential chemical set to approximately 10,000 compounds. Although this step excludes special chemical structures like radicals, it benefits the vast majority of chemical product designs by limiting the results to only structurally stable chemical products.

Information about chemical structure and physical-chemical properties can also be added to the problem using physical property models. After the chemical structures are generated, a set of property filters can be applied to the candidates such that only molecules with appropriate properties remain in the candidate set. This type of procedure is the “test” part of the “generate and test” procedure. One way of testing the candidate list is to match solutions to products that have been previously discovered and validated. In the example, the 10,000 potential candidates

are checked against compounds having exactly 24 atoms in the BioPath.Explore database (Ertl 2010) and results in a significantly reduced candidate list of only 14 options. A second way of testing the candidate list is to identify the candidates that deliver a desired physical-chemical property or attribute. For example, suppose the process or consumer requires that the product have at least a boiling point above 673K. Using the ACD/Labs property prediction tool (de Bie, 2010), boiling points are predicted for the candidate products. Four structures have boiling points below the minimum and are removed resulting in a 10 structure candidate list.

It should be noted that the candidate set used in the example represents all mathematically feasible solutions reduced by a set of heuristics provided by the chemical product designer. The initial list of feasible solutions can be significantly reduced prior to screening by using molecular group theory to only generate chemically feasible products with complete functional groups and group combinations. Functional group theory and the property prediction techniques based on group theory are discussed in more detail in Section 4.2.

Integrating the generation step with the screening step within a mathematical formulation using an objective function is known as optimization. By definition, optimization determines the best or ‘optimum’ product that meets the constraints on the design in the formulation. There are two main methods for finding the best product: multi-criteria decision making (MCDM) and mathematical optimization. In both instances, the optimality of different products is evaluated using formulations with the objective function present (variations 4 or 6). However, the mathematical complexity of optimization procedures can significantly reduce the efficiency of chemical product design problems. In addition, since chemical product designs must be validated with experiments, it follows that the more efficient “generate and test” procedure should be used to screen the candidates and optimization should be used to choose the best

candidate to validate with experiments. A detailed discussion on MCDM and mathematical optimization is available in Appendix A1.

4.2. Useful Property Prediction Techniques

Over the past century, consumer attributes and physical-chemical properties have become an increasingly important element of industrial and engineering chemistry (O'Connell *et al.* 2009). In particular, the uses of physical-chemical properties have become more sophisticated as new chemical products with precise molecular architectures have entered the marketplace. The multitude of competing chemical mechanisms that control these properties has required chemical product designers to become more efficient. Using integrated, computation based design methods based on accurate data, chemical product designers are developing and applying complex models that reveal the conditions needed to attain desired product content and quality (O'Connell *et al.* 2009). These models take many forms, from *ab initio* deterministic models to purely experimental modeling techniques like multivariate linear regression. Regardless of the type of property model used, it is essential that they are accurate, reliable, and computationally efficient throughout the entire design domain. It is also important to recognize that one particular modeling technique should not become the panacea for all problems, rather the “tool driven” approach should be replaced with a “tool box” approach, where the solution method is determined by the features of the problem at hand (Venkatasubramanian *et al.* 2006; Venkatasubramanian 2009).

The roles of property models are intrinsically tied to their use in product design. Since product design may involve both process and molecular engineering, the use of property models will be utilized in different roles depending on the application. Gani and O'Connell (2001) describe three distinctive roles of property models in design: service role, advice role, and the

integration role. In the service role, property models provide estimates where needed for the process simulation models. These process models use two kinds of properties: measurables, such as temperature, pressure, and composition, as well as conceptuels, such as enthalpy, entropy, fugacity, or chemical potential (O'Connell *et al.* 2009). In the advice role, the property models also provide feedback on combinatorial and compositional feasibility, accuracy, and reliability. In the integration role, the property models are reconfigured so that the feasibility constraints can be an integral part of the design process, controlling the design of possible process and product configurations.

Although property models are generally recognized for their traditional service role, the advice and integration roles are actually more powerful (O'Connell *et al.* 2009). These two roles improve design by widening or narrowing the search space when required, thus increasing solution method efficiency (O'Connell *et al.* 2009). However, since the property models may be linear or nonlinear, the forward based advice roles can become computationally cumbersome. For example, the most common model formulation is a generalized expression, often in the form of corresponding states where the variables are scaled by pure component critical properties and combining rules are used for mixtures, as is the case for the Peng-Robinson (PR) equation of state. Although one of the most expansive equations of state in use, it can become tenuous to calculate thermodynamic properties of ternary and larger mixtures, as well as molecules exhibiting long range order. There is also no guarantee that algorithmic programs using the PR will converge to its optimum solution due to its nonlinear nature. Hence, there is a need to balance accuracy and reliability against computational efficiency when identifying the appropriate property models in advice and integration roles. Other requirements, such as dependability, accessibility, generality, and effort will also need to be evaluated (O'Connell *et al.*

2009). This presents a real challenge for chemical product designers with limited knowledge and experience whose goal is to develop quantitative descriptions of natural phenomena that are as varied as the phenomena they describe. In order to maximize the effectiveness of the model and minimize the researcher's effort, a classification scheme was developed by O'Connell *et al.* (2009). Using a similar scheme to that of general product design, property models and/or their parameters are classified as follows: (1) retrieval from computerized databases, (2) estimation via prediction, and (3) new data generation via simulation or experimental measurement (Xin and Whiting 2000).

For Class 1 property models (i.e. retrieval from computerized databases), large amounts of property data have already been gathered by researchers and placed in large databases. Search algorithms can then be used to find models and their parameters. The concern, of course, is that the needed property and molecular architecture will not have been captured, or that an error in consistency, tabulation, and/or omission has occurred (O'Connell *et al.* 2009). To mitigate this risk, Gani and O'Connell (2001) recommend that users should always verify the reliability of vital data by comparisons among sources and by using fundamental variations with state variables as given by thermodynamic equations and derivatives. In most cases, however, the incompleteness of databases makes them best suited for advice roles in chemical product design.

The development of Class 2 property models (i.e. estimation via prediction) is the approach to which chemical engineers are most accustomed. The approach to developing property models and their parameters is shown in Fig. 4.1.

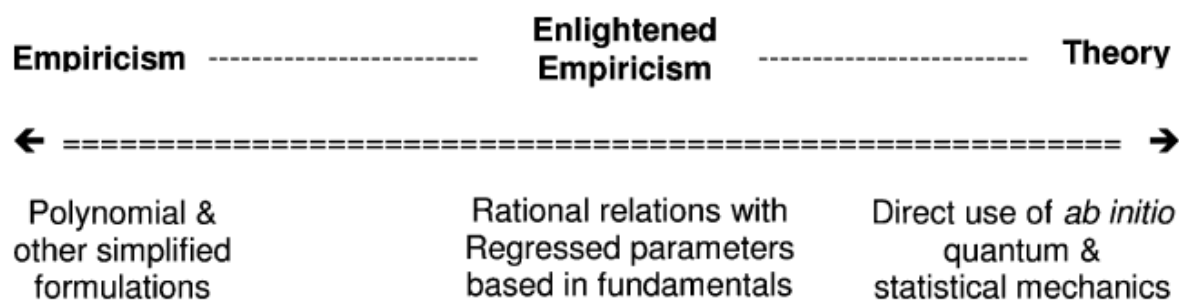


Figure 4.1: Types of Methods to Find Parameters for Property Models (O'Connell *et al.* 2009).

In general, parameters and models exist under different levels of empiricism, moving from easy to use data dependent regressed models whose application range is small, to computationally complex models like *ab initio* which can be applied for any chemical system. In the middle are the most common methods of Group Contribution (GC) and Quantitative Structure Activity Relationships (QSAR) and Quantitative Structure Property Relationships (QSPR) using Topological Indices (TI). These methods balance full empiricism with basic chemical theory to adapt the rigorous equations that describe molecular architecture into a workable form suitable for design.

For Class 3 property models (i.e. new data generation), the controlling mechanisms of the system being studied are not well understood and the trend is to use experimentation or molecular simulation. A more detailed description of property models based on experimentation, specifically the data driven models derived from experimental design, chemoinformatics, and quantitative structure property relationships, is discussed in Chapter 3. An alternative method to experimentation is to use molecular simulation to compute properties using assumed inter- and intramolecular potential functions, or force fields, to estimate a statistical mechanical relationship (O'Connell *et al.* 2009). Used correctly, these methods have the potential to be more cost effective than real experiments. Plus, they can be used to obtain information on systems with molecular architecture that would be otherwise impossible to measure in the laboratory. Of

course, the concern is that the models require stringent application of parameters specific to the system they are modeling, so much that practical application within design has been limited. As such they have been relegated to provide primarily qualitative insights on molecular architecture behavior and their impact on physical and chemical properties. A detailed discussion on simulation based property modeling is available in Appendix A4.

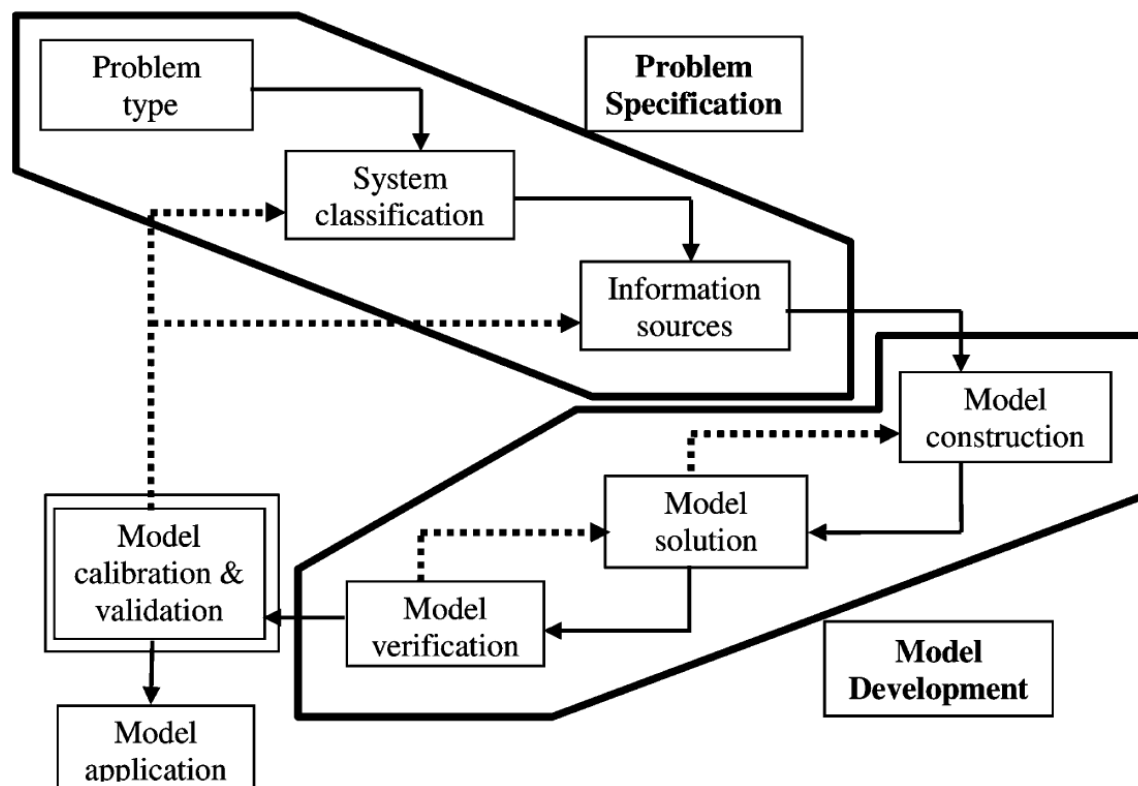


Figure 4.2: Property Model Development Scheme for Chemical Product Design (O'Connell *et al.* 2009).

After the appropriate property model class is chosen, the model is developed to give accurate results efficiently, converging to correct solutions given variable constraints on the molecular architecture. Model robustness must also be developed and sensitivity mitigated in order to avoid situations where parameter uniqueness cannot be established (Gani 2004; O'Connell *et al.* 2009). When no, or only inadequate, models exist O'Connell *et al.* (2009) proposes an efficient development procedure, shown in Fig. 4.2. The first step is to specify the

problem type, system classification, information sources and measures for the expected outcome. Next accuracy requirements are established including their trade-offs with user effort. Then, the model is built, solved, and verified for internal consistency.

Based on the criteria needed for chemical product design (e.g. efficiency and moderate accuracy for screening, etc.) the appropriate physical-chemical property models should be based on easy-to-use molecular construction techniques. Group contribution (GC) and topological indices (TI) are two such techniques. Both types of models utilize quantitative structure property relationships (QSPR) that have been previously developed using empirical relationships between the molecular architecture and the physical-chemical properties found in large databases (Gani 2004).

4.2.1. Group Contribution Methods

Group theory is one way in which a chemical system can be described by a subset of chemical fragments. The fragments are typically a combination of atoms or functional groups with well-described impacts on the overall behavior of the molecule. One of the first methods to quantify the direct relationship of the contributions of group fragments on a molecule's physical properties was the universal functional activity coefficient (UNIFAC) method. It was built for the prediction of activity coefficients directly from atomic and functional group structural units rather than from corrections on molecular solution interaction parameters (Fredenslund *et al.* 1975). For example, in the pure species activity coefficient expression (γ_i^C) as defined in the universal quasichemical (UNIQUAC) method, both the relative molecular volume (r_i) and the relative molecular surface area (q_i) are described by the UNIFAC group contribution terms, R_{gi} and Q_{gi} , respectively:

$$\ln \gamma_i^C = 1 - J_i + \ln J_i - 5q_i \left(1 - \frac{J_i}{H_i} + \ln \frac{J_i}{H_i} \right) \quad (4.10)$$

$$J_i = \frac{r_i}{\sum_w r_w x_w} \quad (4.11)$$

$$H_i = \frac{q_i}{\sum_w q_w x_w} \quad (4.12)$$

$$r_i = \sum_{g_1} n_{g_1 i} R_{g_1} \quad (4.13)$$

$$q_i = \sum_{g_1} n_{g_1 i} Q_{g_1} \quad (4.14)$$

Extensions of this approach to other relevant properties, collectively referred to as Group Contribution Method (GCM), has been the subject of research across many disciplines (Benson 1968; Ambrose 1978; Ambrose 1980; Joback and Reid 1987; Horvath 1992; Constantinou and Gani 1994; Constantinou *et al.* 1995; Marrero and Gani 2001). The culmination of this work has resulted in the current state-of-the-art expression for molecular group contribution shown in Eq. 4.15) (Marrero and Gani 2001). It should be noted that the property contributions of the individual groups are determined by the regression of large data sets. As such, the application of the approach is limited to only similar systems. However, the advantage of this approach is that a relatively few number of groups can be combined to form a large number of molecules within the system of study (Smith *et al.* 1996).

$$P^M = \overbrace{\sum_{g_1} n_{g_1} P_{g_1}}^{1^{\text{st}} \text{ Order}} + \overbrace{w_S \sum_{g_2} n_{g_2} P_{g_2}}^{2^{\text{nd}} \text{ Order}} + \overbrace{w_T \sum_{g_3} n_{g_3} P_{g_3}}^{3^{\text{rd}} \text{ Order}} \quad (4.15)$$

In first order GCM, a molecule is a collection of a number of various simple groups n_{g1} and their respective property contributions P_{g1} that, when combined, predict the properties of the complete molecule P^M . This simplistic relational structure means that the accuracy of the basic group contribution methods is acceptable only in the design of simple monofunctional molecules that do not contain strong group interactions. To increase the accuracy of GCM, further levels of molecular groups have been developed (Constantinou and Gani 1994). Beyond the basic level, known as first order groups g_1 , is the next higher level called second order groups g_2 . They essentially represent different types of interactions among the first order groups and the effects of certain molecular group combinations on the property of the final molecule. As such, second order groups are built from first order groups (shown in Chapter 8) to describe property corrections for interactions P_{g2} where appropriate. w_s represents a property dependent constant that aides in centering the second order property correction. The combined first and second order groups can provide a better description of compounds having many functional groups and/or isomers. However, even the addition of second order groups may not be able to correct for poly-ring compounds and open-chain polyfunctional compounds with more than four carbon atoms in the main chain (Marrero and Gani 2001). Therefore, a further level of molecular groups have been identified and their property contributions have been regressed (Marrero and Gani 2001). The formation of third order groups g_3 is analogous to the second order groups, but the focus is on correcting the property estimate P^M with P_{g3} and w_T for multi-ring compounds, fused ring compounds, and compounds with many functional groups in the structure. As with second order groups, third order groups also have first order groups as their building blocks. The combined expression shown in Eq. 4.15 can represent a significant number of molecular

structures with a good degree of accuracy, making it a useful tool in CAMD algorithms, as discussed in Section 4.4.

Similarly, GCM can also be extended to the prediction of the molecular backbone of polymer systems. This application of GCM is intended to provide a method of estimation and/or prediction of physical-chemical properties of polymers, in the solid, liquid, and dissolved states when experimental values are not available (Van Krevelen 1990; Bicerano 1996). The contributions of the structural and functional groups in the polymer backbone will be different than those for the same groups in different environmental surroundings. Also, the physical-chemical properties of the polymer described by GCM are not only intrinsic to a substance, like properties for small molecules, but are also somewhat described by the size, shape, and morphologies of the nano-, micro-, and meso-structures of the system. To maintain the efficiency benefits of GCM in this situation, a few *fundamental* property descriptors are used to describe a set of *derived* physical-chemical properties based on established modeling interactions (Van Krevelen 1990; Bicerano 1996). For example, one set of correlations uses molecular weight of the repeat unit, length of the repeat unit, van der Waals volume of the repeat unit, cohesive energy, and rotational degrees of freedom of the repeat unit (Seitz 1993) to describe other properties such as the glass transition temperature (T_g) and melting point temperature (T_m), which, in turn, can be used to describe modes of crystallization and morphology (Van Krevelen 1990). More detail can be found in Van Krevelen (1990) and Bicerano (1996).

In order to utilize GCM to its fullest extent, one needs large databases of molecular structure data and a large number of group parameters. In most cases, Marrero and Gani (2001) provide group information for small molecule thermodynamic information and Van Krevelen (1990) and Seitz (1993) have information for polymeric group information. However, situations

will invariably arise that will require descriptions of groups, molecules, or systems beyond the scope of these databases. It is increasingly common for these situations to be handled by new cyberinfrastructure tools that link multiple databases. One new web based GUI, www.chemeo.com, has been specifically created for the construction of group contributions from the shared databases of the Computer Aided Process Engineering Center (CAPEC) at the Technical University of Denmark (d'Anterrosches *et al.* 2005; d'Anterrosches and Gani 2006; d'Anterrosches 2010). Options for building models for various safety and health, energy, phase change, critical, and other base properties are included. This new integrated database approach will likely see refinement in the coming years as new cyber infrastructure initiatives begin to bear fruit.

In spite of its usefulness, GCM has some inherent limitations; the most important of which is the need for a substantial body of experimental data for a given property in order to properly derive reliable group contribution values (Bicerano 1996). An alternative to GCM is to use topological tools to describe a chemical's structure and then relate these topologies to physical and/or chemical properties. For chemical systems, the topology of a system is simply the pattern of interconnections between atoms (Bicerano 1996). Information regarding the (1) total number of atoms, (2) number of each type of atoms, and (3) the bonding between the atoms is utilized to describe the system in a graphical form called a molecular graph (Kier and Hall 1986). The objects in the graph are called vertices and the lines used to connect the objects are called edges. Here, the vertices may be atoms, molecules, molecular groups etc. and the edges may be bonds, interactions etc. (Trinajstić 1992). For simplicity, molecular graphs are generally represented as hydrogen-suppressed graphs where only the molecular skeletons without hydrogen and their bonds are used. Due to the nature of the construction of topological indices,

smaller experimental databases can be utilized with similar prediction accuracy to that of group contribution. A detailed discussion on topological indices can be found in Appendix A4.

4.3. Computer Aided Mixture Design (CAM^bD)

Computer aided mixture design is the application of a computer algorithm to the mathematical formulation of chemical product design. In most cases, the algorithm is a linear program (LP) or nonlinear program (NLP), whose solutions are described in Appendix A1. The inputs to the algorithm are chosen by the researcher. First the chemical constituents used in the design are defined and then a suitable property model is selected whose type (linear or nonlinear) determines the solution strategy the algorithm will use to solve the problem. A suitable mixture range for each chemical constituent is usually specified by the researcher and used to constrain the problem. Various combinations of chemical constituents are then generated and tested against a consumer target property range.

$$P_j^L \leq P_j \leq P_j^U \quad j \in p \quad (4.16)$$

The maximum and minimum of each chemical constituent is then returned as the solution to the design problem.

$$x_i^L \leq x_i \leq x_i^U \quad i \in u \quad (4.17)$$

Obviously, as the number of chemical constituents increases, the solution to the design problem becomes cumbersome. However, the efficiency of this approach can be significantly enhanced by reversing the order of the design by setting property targets first, then substituting a linear mixture model and rearranging to explicitly describe a set of upper limits on the mass fractions as a function of each pure component property.

$$x_{ji}^U \leq \frac{P_j^U}{P_{ji}} \quad (4.18)$$

The minimum of the UL for each i chemical constituent across all properties $j \in p$ is taken as the overall maximum of that chemical constituent in the system.

$$0 \leq x_i \leq x_i^U \quad (4.19)$$

It should be noted that nonlinear property models placed in to Eq. 4.16 result in an implicit description of the component domain that must be solved for x_i^U and is discussed in more detail in Chapter 7.

4.4. Computer Aided Molecular Design (CAMD)

Computer aided molecular design is the application of the generate-and-test procedure outlined in Section 4.1 to the design of molecular chemical products using combinatorial property models like GCM. The forward based solution to the design problem begins by specifying the number and types of groups to be utilized. Next, all potential combinations are generated. The potential combinations are then screened for molecular stability using Eq. 4.9 adapted for GCM:

$$FBN = \sum_{g_1=1}^F n_{g_1} FBN_{g_1} - 2 \cdot \left(\sum_{g_1=1}^F n_{g_1} - 1 \right) - 2 \cdot N_{RG} \quad (4.20)$$

where the FBN is the free molecular bond number of the formulation, n_{g1} is the number of occurrences of group g_1 , FBN_{g1} is the unique free bond number associated with group g_1 , N_{RG} is the number of rings in the formulation, and F is the total number of first order groups. Molecules having a FBN of 0 are then screened using a set of property targets based on consumer or customer preferences.

$$P_j^L \leq P_j^M \leq P_j^U \quad j \in p \quad (4.16)$$

Solutions that meet the constraints of Eq. 4.16 are deemed candidate solutions.

The efficiency of this approach can be significantly enhanced by reversing the order of the design by setting property targets first, then substituting a first order GCM model and rearranging to explicitly describe a set of upper and lower limits on the number of similar groups for each property.

$$n_{jg_1}^U \leq \text{int} \left(\frac{P_j^U}{P_{jg_1}} \right) \quad (4.21)$$

$$n_{g_1}^L \geq \text{int} \left(\frac{P^L}{P_{g_1}} \right) \quad (4.22)$$

$$n_{g_1}^L \leq n_{g_1} \leq n_{g_1}^U \quad (4.23)$$

The minimum of the upper limit for each group across all properties is taken as the overall maximum number of groups of that group type in any molecule. The inverse does not hold, since it is not a requirement that all dissimilar groups be included in the molecule. Hence, the overall minimum number of groups of a specific group type will always be set to 0. For higher order property description, the combination of Eq. 4.15 and Eq. 4.16 results in an implicit model that must be solved for n_{g1}^U and is discussed in more detail in Chapters 6 and 8. The procedure then continues by building combinations and testing the *FBN*. This type of approach is known as reverse property prediction and is a partial component of the reverse problem formulation (RPF) that is discussed in Chapter 5 and Chapter 6.

4.5. Summary

Although the information presented in this chapter was not a complete description of property models, it nevertheless gives a good overview of what challenges exist. Property models can be derived from experimental data, computed from fundamental chemistry concepts, or as a mixture of the two. It is increasingly clear that both the methods and applications of property models are continuing to change and, as such, researchers must adapt to their use. In particular, alternative modeling methodologies capable of making use of these new property models using knowledge based programming and decision schemes will be paramount (Venkatasubramanian *et al.* 2006; Venkatasubramanian 2009). A great opportunity exists within the discipline to venture into this domain, although, to do so, means overcoming the high barrier of entry regarding the mastery of knowledge representation and search techniques, algorithm engineering, databases, and certain misconceptions about what the intellectual challenges are (Venkatasubramanian 2009).

Chapter 5

Motivation and Challenges

In the preceding chapters, it has been shown that the physical-chemical properties and attributes of chemical products are controlled by a multitude of separate and often competing mechanisms operating over a wide range of length and time scales. Quantifying these behaviors and incorporating them within the chemical product design problem has generally been performed on a case-by-case basis using either experimental or simulation techniques. The inefficiency of these methods has limited researchers to screening only a handful of products prior to product and process development, posing great economic risk of designing non-optimum products. Finding a more efficient approach that is capable of screening large numbers of chemical products prior to development is the focus of this dissertation. As suggested by the National Research Council's Committee on Integrated Computational Materials Engineering (NRCCICME) this can be achieved through the integration of the process, product, and property models utilized in design (Committee on Integrated Computational Materials Engineering 2008). Because of the immense data structure, number and types of models, and the attribute and physical-chemical properties utilized, such integration is a "Grand Challenge" (Committee on Integrated Computational Materials Engineering 2008).

To achieve integration, computational models used in product design needed to be closely coupled with manufacturing and distribution techniques that exist at multiple scales. The process systems research community has approached this problem by shifting focus from the

design of the physical form, function, and aesthetics of assembled products (e.g. mixtures of molecular architecture) to the design of chemically *formulated* products (e.g. a single molecular architecture) (Hill 2004). Chemically formulated products are products designed at the molecular level to deliver a specific desired attribute that may exist at multiples scales such as the quantum-scale, nano-scale, meso-scale, macro-scale, and mega-scale. Examples of formulated products include pharmaceuticals (Solvason *et al.* 2009), bio-inspired devices (McAllister and Floudas 2008), semi-conductors (Chalivendra *et al.* 2009), processed foods (Hill 2004), personal hygiene products (Hill 2004), and many more product types. Current state-of-the-art product design, or formulation, uses a generate-and-test approach, where the ‘generation’ is experimentation, prediction, or simulation and the ‘testing’ is a validation that the results of the generation meet the target criteria. With the recent improvements in computational power, simulation is increasingly becoming a viable tool in the generation step because of its wide application range. By using a small number of parameter estimates, simulation techniques compute molecular architecture information at the smallest scale, passing it to progressively larger length scales, under high accuracy as shown in Fig. 5.1. However, each simulation can potentially take weeks to perform. For instance, when determining micro-scale dependent property values such as polymer permeation, Satyanarayana *et al.* (2009) utilized a united-atom MD simulation of polyisobutylene under relaxed configurations based on models developed by Tsolou *et al.* (2008) that took six weeks to calculate using a parallel program on a super computer. Since, C₃₂₀ polyisobutylene was one of 12 possible polyolefins identified by the Van Krevelen (1990) group contribution method, it could conceivably take the better part of a year to find a suitable polymer for this portion of the chemical product design, which is far from ideal.

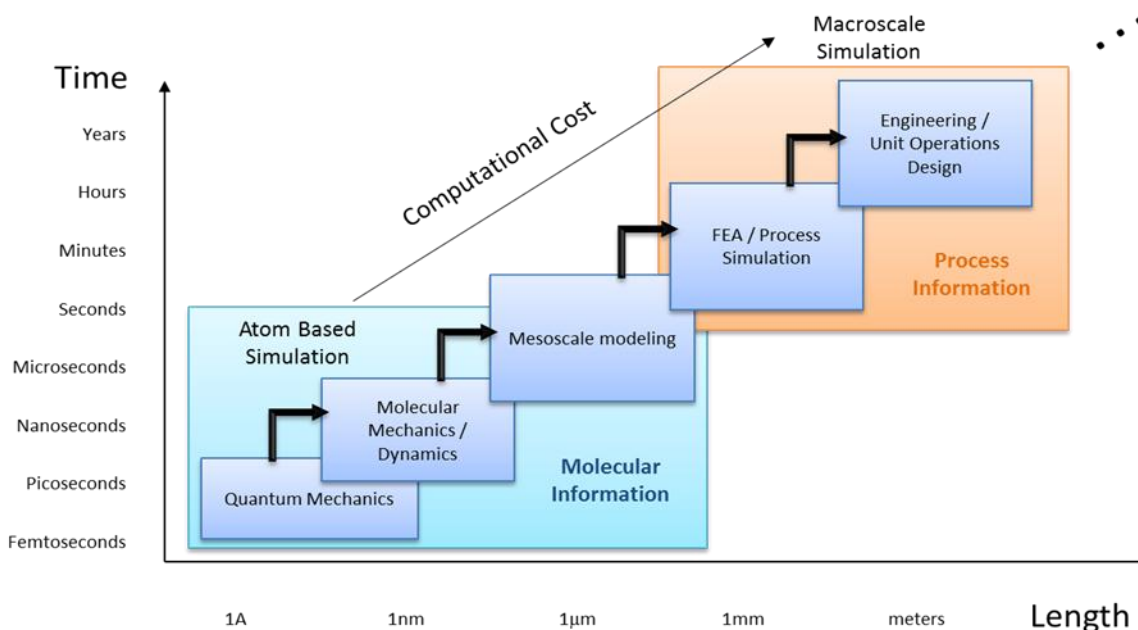


Figure 5.1: The Computational Tools used in Chemical Product Design for the Transfer of Information Between Multiple Scales using Nested Simulation.

Alternatively, experimentation could be utilized, but the methods for making and testing new polymers can be just as time consuming as simulation in addition to containing large amounts of uncertainty. For example, as discussed in Chapter 3, it is possible to utilize experimental characterization data to describe the molecular architecture at each scale after the various molecular level architectures have been identified. This approach requires the researcher to create the chemical product, providing instant validation of the designed structures and manufacturing processes. Thus in order to capture the exact polyisobutylene structure Satyanarayana identified, an appropriately sized system would need to be suspended in the correct solvent, molecularly characterized using nuclear magnetic resonance (NMR), polymerized to the appropriate size, and tested for the necessary microstructure using scanning electron microscopy (SEM). If any of these steps cannot be performed, it can lead researchers to prematurely dismiss an optimum chemical product.

Finding a way to link experimental and simulation approaches would offer some significant advantages in chemical product design. For instance the targeting of potential breakthroughs using enhanced algorithms would reduce the number experimental or simulation design procedures required for validation (Fermeglia and Pricl 2009). Furthermore, the ability to bridge domains via models could potentially allow nested, nonlinear routines to be reduced to a single sub-problem. Discovering the appropriate linkage is one of the keys to the methods developed in this dissertation.

5.1. Current Limitations of Chemical Product Design

The traditional method of solving a chemical product design problem has been to compute information at smaller length scales and pass it to models at larger length scales. In this approach the linkage of the scales would be critical as would be the choice of basis set from which the model parameters are estimated. While often the most accurate method for predicting properties, simulation has two limitations: (1) it has an immense computational cost due to hierarchical nesting and (2) it utilizes *a priori* knowledge of the molecular architecture (i.e. the number and types of atoms or electrons present). The large computational cost typically prevents an accurate modeling of mesoscopic structure such as the morphology of polymers without the use of constraints that significantly limit the degrees of freedom in the simulation (Fermeglia and Pricl 2009). Furthermore, when this method is integrated within the product-process design framework (i.e. using a simulation model as Eq. 4.4) the computational intensiveness exponentially increases yet again since each projected molecular architecture must be simulated to determine its physical-chemical properties. To minimize the computational cost in these types of problems, the property prediction simulations are typically approximated with constitutive equations based on structure descriptor models such as property based group

contribution method (GCM), quantitative structure property/activity relationships (QSPR/QSAR) using topological indices (TI), or chemometric based models as shown in Fig. 5.2. As shown in pathway “b” in Fig. 5.2, the introduction of structure descriptor models discussed in Chapter 4, improves computational efficiency by avoiding nested simulation at the cost of prediction accuracy. However, these descriptor models are still computationally expensive when utilized in design algorithms because of their often highly nonlinear nature.

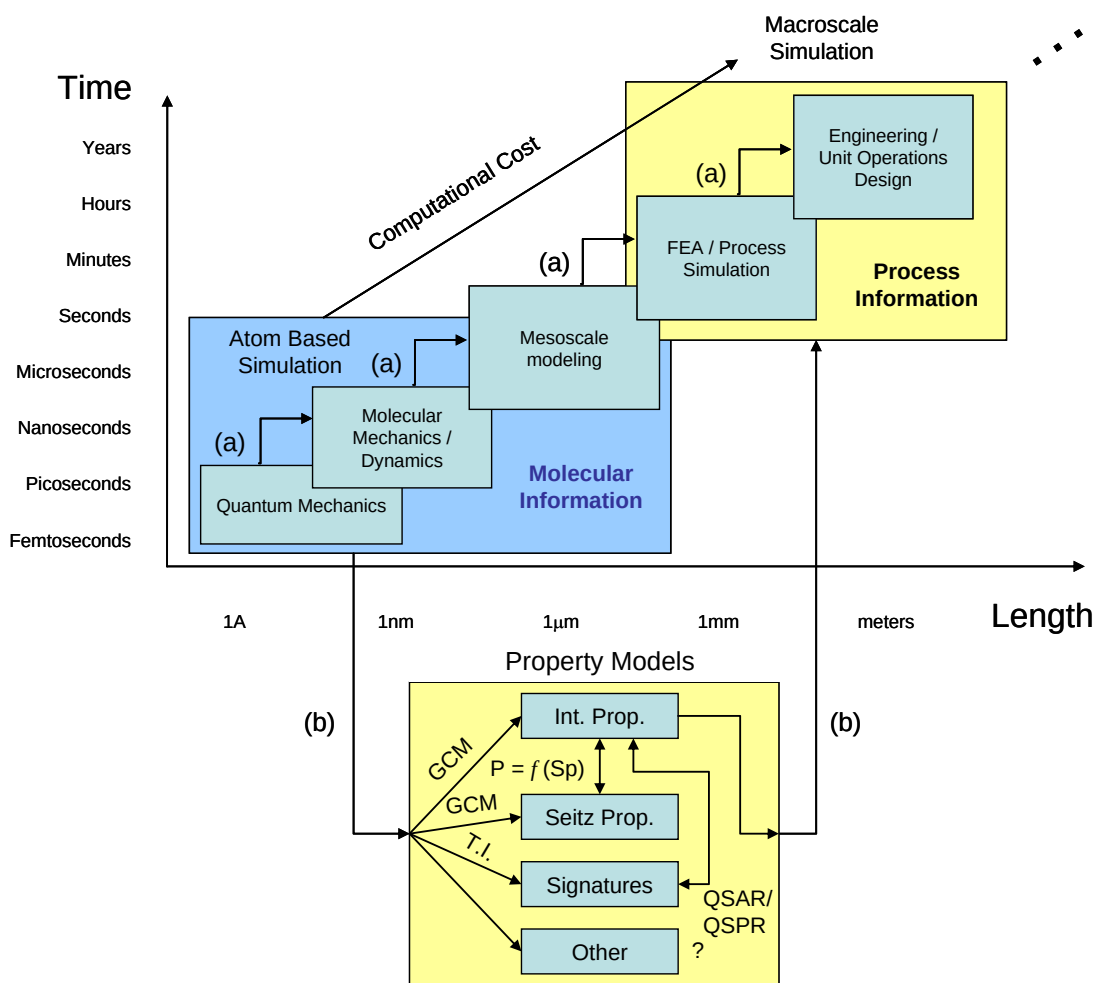


Figure 5.2: The Computational Tools used in Chemical Product Design for the Transfer of Information Between Multiple Scales using: (a) Nested Simulation and (b) Structure Descriptor Models.

The second major limitation of the current design methodologies is the inability to accurately predict attribute responses as a function of molecular architecture. Consumer attributes are a notoriously difficult set of properties to define, and, as such, rely heavily on empirical correlative and causal relationships based on experimental study. Since a product's consumer attributes are what consumers and customers use to judge the success of a chemical product, they serve as the ultimate validation step. In the NRCCICME (2008) published report, it was suggested that the best approach to multiscale product design was to utilize complimentary experimental and theoretical based models since no model, as of yet, has shown the ability to operate accurately across the diversity of length scales (Committee on Integrated Computational Materials Engineering 2008). Since Hill (2004, 2009) further states that the goal of the PSE community is to develop computational product design tools to guide and focus experimentation; it follows that the relationship between the underlying fundamental physical-chemical properties, molecular architecture, and consumer attributes will involve empirical relationships. Some authors, such as Smith and Ierapetritou (2009), formulated a chemical product design problem using consumer attribute derived models as a function of consumer preference tests, as shown in Fig. 5.3.

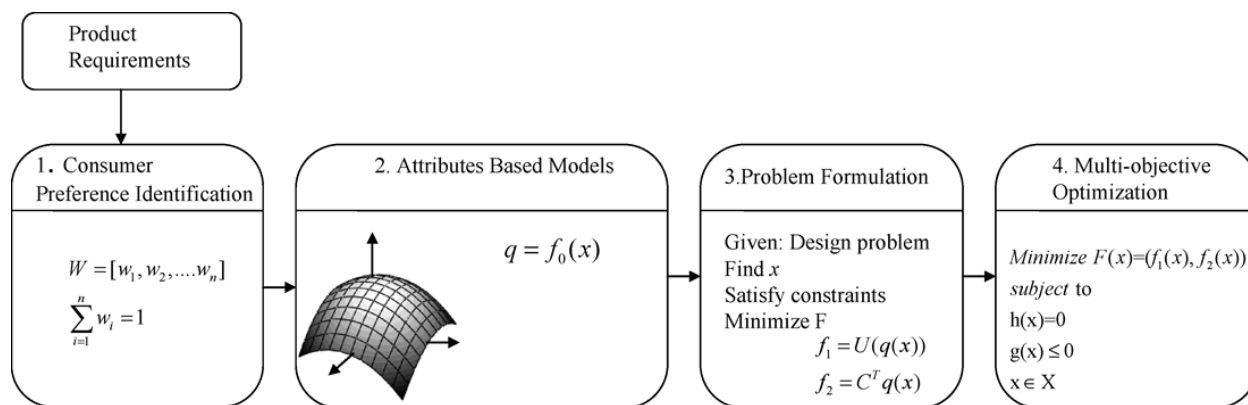


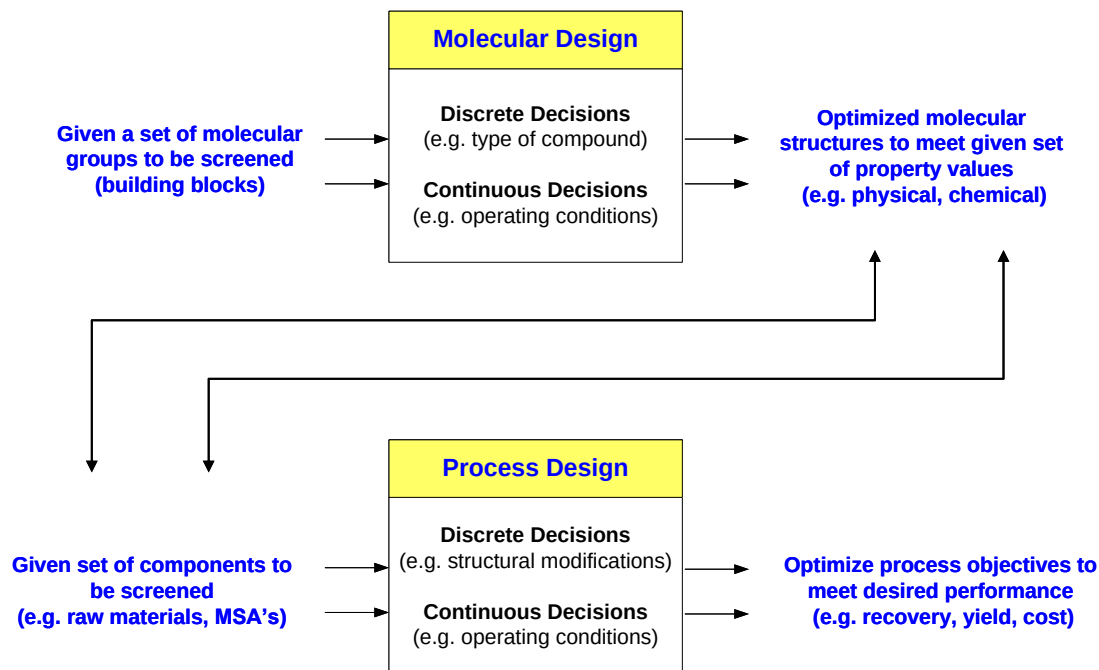
Figure 5.3: Schematic of Chemical Product Design using Consumer Preference (Smith and Ierapetritou, 2009)

In this approach, the consumer attribute models are derived as needed from the consumer preference and chemical constituent databases. The attribute models are then utilized as additional property models in the mathematical formulation (i.e. Eq. 4.4). As with the database approaches discussed in Chapter 3, the biggest limitation of this design methodology is the inability to conduct true formulation design; only products and mixtures of products already in the database can be solutions to the design problem. An alternative formulation that circumvents this limitation is presented in Chapter 8. It utilizes GCM models and attribute-property relationships to design chemical products without the use of a molecular architecture database. The capability of the formulation is extended in Chapter 9 to handle characterization-based group contribution models (cGCM) that are capable of describing a larger variety of molecular architectures.

5.2. Reverse Problem Formulation

The focus of the dissertation is on the development of efficient design methods and techniques for multiscale chemical product design based on consumer attributes. One of the tools recently developed by Eden *et al.* (2003) is employed to assist in this endeavor: the reverse problem formulation (RPF). The reverse problem formulation (RPF) is a procedure for reconfiguring the generate-and-test design strategy into a more computationally efficient scheme (Gani and Pistikopoulos 2002, Eden *et al.* 2003, Gani 2004). Using the duality of linear programming, the RPF solves the design problem in the lower dimensional physical-chemical property and/or consumer attribute domain first, and then generates only the components that have properties that match the solution. For example, in a conventional product design problem, candidate molecules are first generated from a set of structural features in a molecular design and then fed as inputs to a process design simulation as shown in Fig. 5.4a.

(a)



(b)

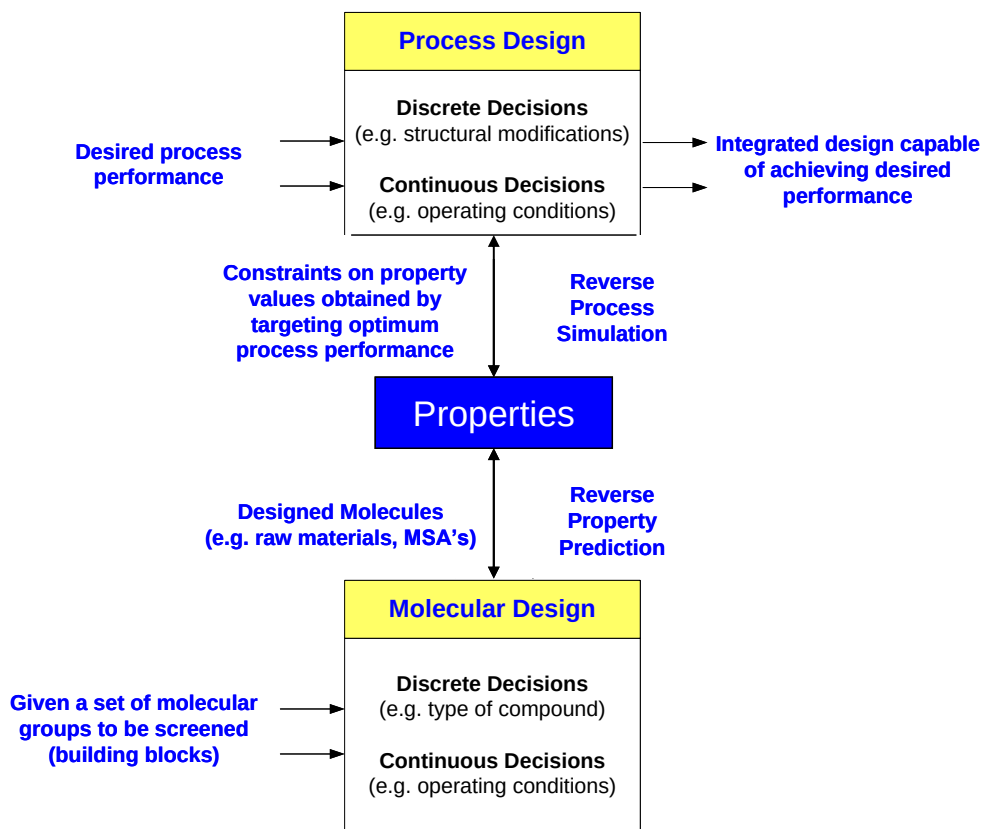


Figure 5.4: A Product and Process Design Problem using (a) a Sequential Solution Format and (b) a Reverse Problem Format (Simultaneous) (Eden *et al.* 2004; Gani 2004; Eljack *et al.* 2007).

In most cases, the majority of candidate solutions fail to meet the test criteria and are screened out in either the property prediction or process simulation steps. This is a highly inefficient process since resources are wasted calculating non-target solutions.

As shown in Fig. 5.4b (i.e. a RPF), target physical-chemical properties or attributes can be determined from the solution of the reverse simulation problem and, as long as these targets are matched, any number of property models may be used to generate the molecular architectures at the various scales to ensure a solution (Eden *et al.* 2004). This design structure allows the RPF to efficiently handle complex MINLP formulations that utilize non-linear property models without requiring tedious solution strategies.

5.3. Challenges Addressed in this Dissertation

One of the goals of this dissertation is to demonstrate how the RPF can be used to bridge the length domains shown in Fig. 5.1 by creating insightful shortcuts that allow macro-scale constraint information to be mapped down to the appropriate molecular architecture scale. This new source of information can then be used to reduce the number of potential candidates evaluated by moving the testing or validation function from the macroscale down to the smaller molecular architecture scale, prior to computation. A visual description of this method is shown in Fig. 5.5. The traditional solution approach sends information peripherally through the length scales in a nested simulation or experimentation approach (i.e. pathway “a”); whereas the goal of this work is to link the scales with a bridge through a ‘property domain’ that will allow for a much more efficient solution that will significantly reduce the computational time of multi-scale design problems (i.e. pathway “b”).

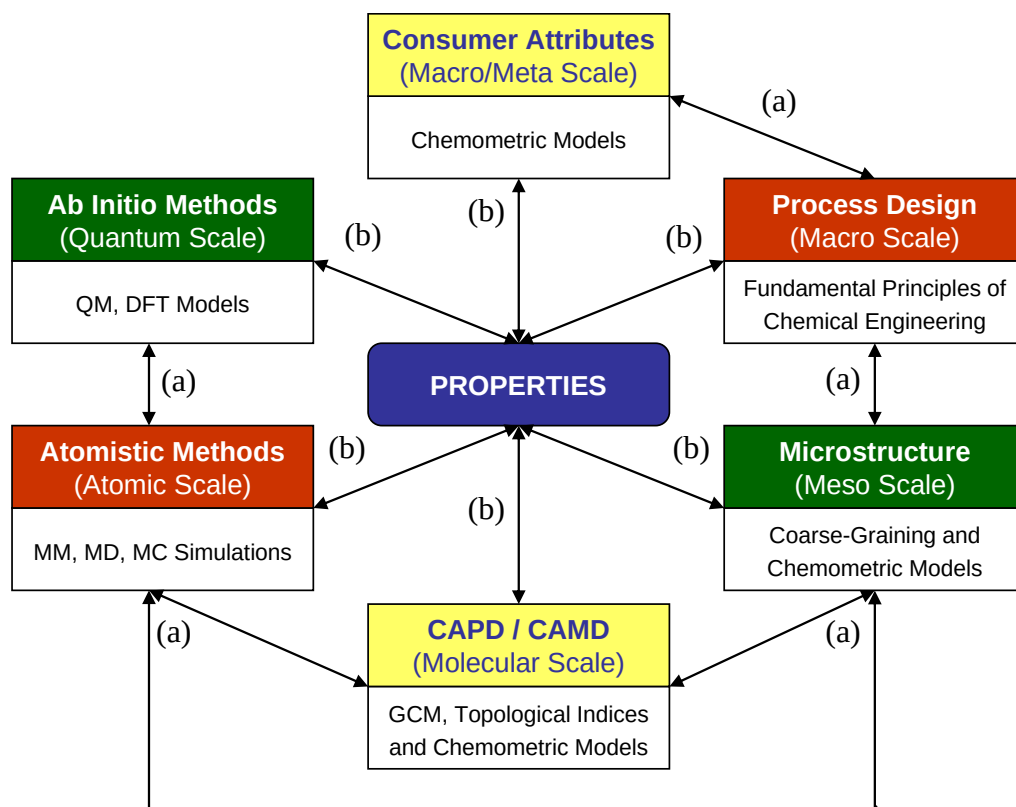


Figure 5.5: A Multiscale Product Design Framework Showing (a) the Traditional Approach to Computational design and (b) a RPF Approach Linking each of the Scales via a Common Property Domain.

The major limitation in achieving this goal is finding the appropriate property domain in which to decouple the constitutive equations describing the molecular architecture from each other. Many potential property domains exist. For instance, the computational chemistry methods described in Appendix A4 were developed to model the potential energies of individual molecules and interaction forces between molecules. In contrast, most engineering design problems are based on chemical and material constituents that have physical-chemical properties that can be represented as a continuum. Developing a universal link between the two methods has been the focus of thermodynamics and statistical mechanics whose parameters may represent another potential property domain. Furthermore, the materials that have the biggest potential for transforming our society have properties dependent on nanostructure and microstructure in the

meso-scale domain. The molecular architecture at this length scale is too large to be computationally designed solely with molecular simulations like molecular mechanics (MM), molecular dynamics (MD), and monte carlo (MC), resulting in a trend toward estimations via hybrid models and uncommon property domains. Some preliminary studies have shown that sequential approaches using reverse problem formulations (RPF) on multiple property domains offer computational advantages. For example, structure based property models like group contribution method (GCM) (Marrero and Gani 2001), GCM plus connectivity indices (GC+) (Chemmanurgattuvalappil *et al.* 2009; Chemmanurgattuvalappil *et al.* 2010), and signature descriptor based TI methods (Chemmanurgattuvalappil *et al.* 2009; Chemmanurgattuvalappil *et al.* 2010) have expanded the number and types of properties they can describe, which allows them to replace a number of simulation models. However, accuracy limitations will constrain these approaches to short range order applications (e.g. less than 60-100 atoms). To address longer range order, a partial RPF has recently been combined with grid technology for polymer design by Satyanarayana *et al.* (2009). As discussed in the polyisobutylene example, this approach uses inverse GCM models to target monomer precursors with desirable properties followed by simulation of the polymer structures to evaluate their meso-scale impacts. Although an improvement in design at the meso-scale, the method is not simultaneous, and is therefore computationally inefficient. As a result the method has been limited to simple polymer structure investigations and is unable to address more complex nano- and micro-structure arrangements. In conclusion, finding a suitable property domain to perform the design remains case-specific and subject to the researchers desire to balance efficiency vs. accuracy. Improving the objectivity of this approach will be necessary to improve multiscale chemical product design.

Another goal of this dissertation is to develop a way to utilize experimental data, parameters, and models in a RPF. The reason behind choosing to work with experimental models comes from the nature of the intended end-use of chemical product design algorithms. Since the effectiveness of a chemical product is most often determined by its consumer attributes, and since the consumer attributes are most likely quantified using data from consumer preference tests, then it follows that any framework should at least be partially based on experimental data. This observation does not preclude using the more common physical and chemical properties found in process design in the framework, rather, it only means that experimental data needs to be integrated into the design in some fashion. Furthermore, while the use of data places constraints on the application range of the tools developed for multiscale product design, it does not affect the fundamentals of the methods developed and could be applied to systems described by deterministic relationships (e.g. computational chemistry methods) if desired.

In this dissertation, three methods for improving the solution strategy for solving multiscale chemical product design problems are developed. All three methods use the newly developed property clustering (PC) algorithm (Shelley and El-Halwagi 2000; Eden *et al.* 2004). Property Clustering (PC) is a mathematical technique that decomposes the problem as a Euclidean vector consisting of unit property clusters that describe orientation of the property in the domain space and a scalar called the augmented property index (*AUP*) that describes its magnitude. A detailed discussion on the benefits of PC and its utilization in design algorithms is provided in Chapter 6.

Attribute-computer aided mixture design (aCAM^bD) is the first of the three methods developed to address the current limitations in multiscale chemical product design. It recognizes

that through the use of transformation functions, experimental design techniques and data-driven property/attribute models can be directly utilized in the PC algorithm to solve chemical product design problems using a RPF. Approaching the design problem in this manner significantly reduces the computational complexity of the problem by using the duality of linear programming to solve the design problem in the lower dimensional property or attribute domain instead of the high dimensional component space. The method works by first solving for a target attribute or physical-chemical property region and then generating solutions using predictive models based on design of experiments (DOE) that have attributes that fall within that region. Furthermore, for design of three properties, the design space and subsequent solution can be visualized on a single diagram, regardless of the number of chemical constituents in the design, which provides insight into the nature of the problem. A detailed discussion of this method, including how it is built to utilize data driven property models, regressor based property clusters, and the reverse problem formulation, as well as a case study on a spun yarn polymer blending is discussed in Chapter 7.

Recognizing that aCAM^bD is limited to only mixture designs, and that the effectiveness of most chemical products is determined by its consumer attributes, which seldom have adequate quantitative structure-property (QSPR) or quantitative structure-activity relationships (QSAR), a second new method was developed to map consumer attribute data into a more conventional physical property domain where group contribution methods (GCM) are used to find new chemical structures. By mapping attribute information to a physical-chemical property domain with strong QSAR/QSPR description, multiple sources of scientific knowledge, information, and DOE data can be combined to design chemical products without the use of a database or *a priori* knowledge of the molecular architecture. The method specifically uses property clustering (PC)

in a reverse problem formulation (RPF) to reduce the computational complexity of the problem and provides insight into the nature of the relationships between molecular architecture and consumer attributes. The method is described in more detail in Chapter 8 where it is applied to the design environmentally benign refrigerants.

A third opportunity for improving multiscale chemical product design comes from the recognition that the descriptive data used in chemical product design is length dependent; each length scale has specific molecular architectures that dominate the chemistry at that scale. Quantification of these architectures using characterization techniques results in large, highly correlated data sets that describe the products molecular architecture. Rather than solving the molecular design using this high dimensional data set, as is done using nested simulation and experimentation, decomposition techniques can be applied to the correlated data set to find a lower dimensional latent variable domain more conducive to chemical product design. The relationship between the latent variable domain and the consumer attributes is then estimated using chemometric techniques (e.g. MLR, PLS, etc.). To perform a chemical product design in this domain requires the creation of a new characterization based group contribution method (cGCM) that can take advantage of the additional molecular architecture information provided by characterization. Information from the molecular scale on short range order, such as group structure, conformation, and stereoregularity, is combined with information from the micro-scale on long range order, such as the size, shape, and aspect ratio of particles within a solid matrix within a single latent variable domain, which drastically improves the computational efficiency of multiscale chemical product design. As demonstrated in Chapter 9, the characterization based computer aided molecular design (cCAMD) method is built specifically to utilize IR/NIR spectroscopy data, PCA decomposition techniques, MLR attribute-latent variable relationships,

latent variable property clusters, and the reverse problem formulation and results in a framework that decouples multiscale design problems from their respective scales, solving them in a low dimensional, computationally efficient manner. A constrained pharmaceutical excipient design for direct compressed acetaminophen tablets is used to illustrate the method.

In conclusion, it is shown in Chapter 7, Chapter 8, and Chapter 9 how a hybrid approach that combines Class 2 models (e.g. GCM, etc.) with experimental data generation and decomposition techniques (e.g. DOE, IR/NIR spectroscopy, PCR, and PLS, etc.) can dramatically improve the computational efficiency of multiscale chemical product design problems. This is a central result of the multiscale chemical product design framework presented in this dissertation. Many opportunities exist for extending this work to include other characterization tools and process design and configuration parameters to develop a truly unique and powerful design methodology within the PSE community.

Chapter 6

Methods for Multiscale Chemical Product Design

One of the most difficult tasks in multi-scale chemical product design is to balance the efficiency of experimental methods against the accuracy of simulation techniques in the computer-aided mixture and molecular designs. It was shown in Chapter 5 that the reverse problem formulation (RPF) is particularly adept at achieving this goal by decomposing the computational difficulty of the design problem into two sub-problems, reverse simulation and reverse property prediction, and matching their solutions to achieve results for the chemical product design (Eden *et al.* 2003). Another tool that aides in the solution of multiscale chemical product design problems is property clustering (PC). Property clustering converts the attribute and physical-chemical properties into conserved surrogate attribute or property clusters capable of describing the design space under reduced dimensionality. Furthermore, property clustering provides excellent visualization of the decomposition and algorithmic solution processes and provides insights into the effects of model parameters on the design. The PC methods for solving mixture and molecular design problems form the basis for the three methods developed in this dissertation.

6.1. Property Clustering Algorithms

Property clustering was developed as a tool for tracking properties in a conserved manner by Shelley and El-Halwagi (2000). It was later applied to process and product design by Eden *et*

al. (2003), Eljack *et al.* (2007), and Solvason *et al.* (2008). As a design tool, the technique uses the RPF to challenge conventional design by solving the design problem in a property cluster domain rather than the chemical constituent domain. The conserved surrogate property clusters are generated from the attributes or physical-chemical properties important to the chemical product. They are described by property operators, which have linear mixing rules, even if the operators themselves are nonlinear. In equation Eq. 6.1, the property, $y = P$, is described by a linear property operator expression:

$$\psi_j(y_j)_M = \sum_{i=1}^u \psi_j(y_j)_i \cdot x_i \quad (6.1)$$

where $\psi_j(y_j)_i$ is the property operator of the pure component property response y_j for chemical constituent i , x_i is the mass (or mole) fraction of constituent i , and $\psi_j(y_j)_M$ is the property operator of the mixture. In many cases the property operator is simply the linear expression of a common physical-chemical property. For example, if the property operator describes density, then to meet the linear criteria imposed by equation Eq. 6.1 we would use specific volume as the property operator, ignoring any interaction effects from mixing, as shown in Eq. 6.2 and 6.3.

$$\psi_j(y_j)_i = \frac{1}{\rho_i} \quad (6.2)$$

$$\frac{1}{\rho_M} = \sum_{i=1}^u \frac{1}{\rho_i} \cdot x_i \quad (6.3)$$

A list of physical-chemical properties commonly used in chemical product design are given in Table 6.1 (Marrero and Gani 2001). Each universal parameter (e.g. T_{m0} , T_{b0} , T_{c0} , P_{c1} , P_{c2} , V_{c0} , G_{f0} , H_{f0} , H_{v0} , H_{fus0}) and molecule specific parameters (e.g. T_m , T_b , T_c , P_c , V_c , G_f , H_f , H_v , H_{fus}) was derived from the ICAS property database (Marrero and Gani 2001).

Table 6.1: Common Physical-Chemical Properties used in Chemical Product Design

Property (y)	Property Operator ($\psi(y)$)
Normal Melting Point (T_m)	$\exp(T_m/T_{m0})$
Normal Boiling Point (T_b)	$\exp(T_b/T_{b0})$
Critical Temperature (T_c)	$\exp(T_c/T_{c0})$
Critical Pressure (P_c)	$(P_c - P_{c1})^{-0.5} - P_{c2}$
Critical Volume (V_c)	$V_c - V_{c0}$
Standard Gibbs Energy at 298K (G_f)	$G_f - G_{f0}$
Standard Enthalpy of Formation at 298K (H_f)	$H_f - H_{f0}$
Standard Enthalpy of Vaporization at 298K (H_v)	$H_v - H_{v0}$
Standard Enthalpy of Fusion (H_{fus})	$H_{fus} - H_{fus0}$

It is often difficult to find adequate linear property operator expressions or models for the attributes and physical-chemical properties used in the chemical product design and the experimenter must search for linearization functions like those proposed by Johnson and Wichern (2007) and Solvason *et al.* (2009) or use an approximation for which the benefits of the PC+RPF method versus the loss of solution certainty must be weighed (Solvason *et al.* 2008). This result often leads the experimenter to limit the application of this method to high volume screening designs (Gate 1 in Fig. 2.5).

After the property operator equations are defined, the method is universal. As shown in Eq. 6.4, the property operators are non-dimensionalized by dividing by a reference property operator (Eden *et al.* 2003).

$$\Omega_{ji} = \frac{\psi_j(y_j)_i}{\psi_j(y_j)_{ref}} \quad (6.4)$$

This step serves the purpose of scaling the property design space to facilitate a better graphical understanding of the problem or, as shown later in Chapter 7, is used to ensure a solution when the property operator equations contain negative parameter estimates. The non-dimensionalized properties are then summed into the Augmented Property Index (*AUP*) (Shelley and El-Halwagi 2000).

$$AUP_i = \sum_{j=1}^p \Omega_{ji} \quad (6.5)$$

A cluster is then defined by dividing the non-dimensionalized property by the *AUP*, as shown in Eq. 6.6.

$$C_{ji} = \frac{\Omega_{ji}}{AUP_i} \quad (6.6)$$

By definition, property clusters must vary between 0 and 1 (Shelley and El-Halwagi 2000; Eden *et al.* 2003). This means that the cluster domain is capable of describing *all* possible property values associated with a chemical product design, independent of molecular architecture. For instance, the internal energy of an arbitrary system can be estimated from a path integral Monte Carlo summation of molecular energies. Using PC, both the molecular energies of a unit cell and the macroscale internal energy can be represented using the same cluster domain. Shown in Fig 6.1 is a ternary diagram, or simplex, of the cluster domain (C_1 , C_2 , and C_3) of three properties (P_1 , P_2 , and P_3) that have been identified as key to the design of a chemical product. Each vertex represents a single property cluster having the value of 1 meaning that the domain constrained by the boundaries of the cluster diagram is the entire range of all current and future molecular

architectures. As a result, the property cluster domain can be used to independently and efficiently evaluate a product’s molecular architecture at multiple scales. This concept is explored in more detail in Chapter 9.

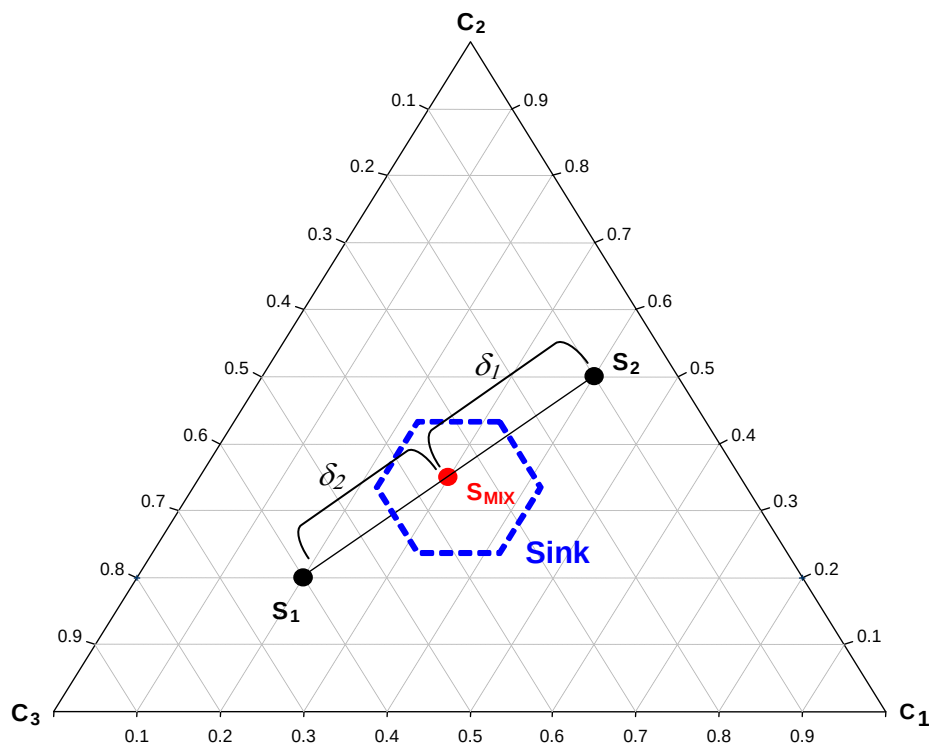


Figure 6.1: A Property Cluster Simplex Diagram with Property Clusters at the Vertices, Pure Component Effects Represented as Independent, Discrete Stream Points, and the Product Targets Represented as a Sink Region (Eden *et al.* 2003).

Also shown in Fig. 6.1 are the pure component property clusters of two chemical products, S_1 and S_2 , as well as a feasibility region, denoted as “sink.” The pure component clusters are calculated from three property values, P_1 , P_2 , and P_3 , of each chemical product (e.g. S_1 and S_2) using Eq. 6.2, 6.4, and 6.5. The feasibility region is a result of the chemical product’s target specifications, defined by upper (U) and lower (L) limits on the properties, and is described by six unique clusters that maximize its size. If a chemical product falls outside the feasibility region then it is not a valid solution to the chemical product design problem. The

feasibility clusters are found by solving Eq. 6.5 and 6.6 using the following dimensionless properties (Eden *et al.* 2003):

$$\begin{aligned} \text{Pt 1: } (\Omega_1^L, \Omega_2^L, \Omega_3^U); \text{ Pt 2: } (\Omega_1^L, \Omega_2^U, \Omega_3^U); \text{ Pt 3: } (\Omega_1^L, \Omega_2^U, \Omega_3^L); \\ \text{Pt 4: } (\Omega_1^U, \Omega_2^U, \Omega_3^L); \text{ Pt 5: } (\Omega_1^U, \Omega_2^L, \Omega_3^L); \text{ Pt 6: } (\Omega_1^U, \Omega_2^L, \Omega_3^U) \end{aligned} \quad (6.7)$$

It should be noted that for each new property added to the design, two new feasibility cluster points are added to describe the feasibility region.

In some situations, the property clusters of the feasibility region or the chemical constituents are negative or difficult to visualize. This is a result of the parameters used in the linearized property models and is often difficult to avoid. To handle this situation, a property reference algorithm was written to find the set of reference property operators that give the largest values of the *AUP* while maintaining positive estimates for the pure component values:

$$\text{Max } \sum_{i=1}^u AUP_i \quad (6.8)$$

$$\text{s.t. } AUP_i > 0 \quad i \in u \quad (6.9)$$

$$AUP_i < M_x \quad i \in u \quad (6.10)$$

$$\psi_j(y_j)_{ref} \neq 0 \quad j \in p \quad (6.11)$$

$$\Omega_{ji} = \frac{\psi_j(y_j)_i}{\psi_j(y_j)_{ref}} \quad (6.12)$$

$$AUP_i = \sum_{j=1}^p \Omega_{ji} \quad (6.13)$$

A more detailed analysis of the reference property algorithm is available in Chapter 7. The procedure for converting a set of chemical product design equations and parameters for use in a computer aided mixture design CAM^bD with property clusters is presented in Table 6.2.

Table 6.2. The CAM^bD Property Clustering Conversion Algorithm

Step	Description	Equation
1	Linearize Property Prediction Models	6.1
2	Guess Property Operator Reference Values and Calculate Pure Component Property Operator Values	6.4
3	Convert UL and LL Feasibility Constraints to Property Operators	6.7
4	Calculate the AUP of the Pure Component Property Operators and the Feasibility Constraint Property Operators	6.5, 6.7
5	Calculate the Pure Component Clusters	6.6
6	Calculate the Feasibility Region Sink	6.6, 6.7
7	Run Reference Optimization Algorithm for Non-Positive Clusters	6.8-6.12
8	Convert all Clusters to Cartesian Coordinates and Plot	6.23-6.26

CAM^bD Property Clustering Conversion Algorithm

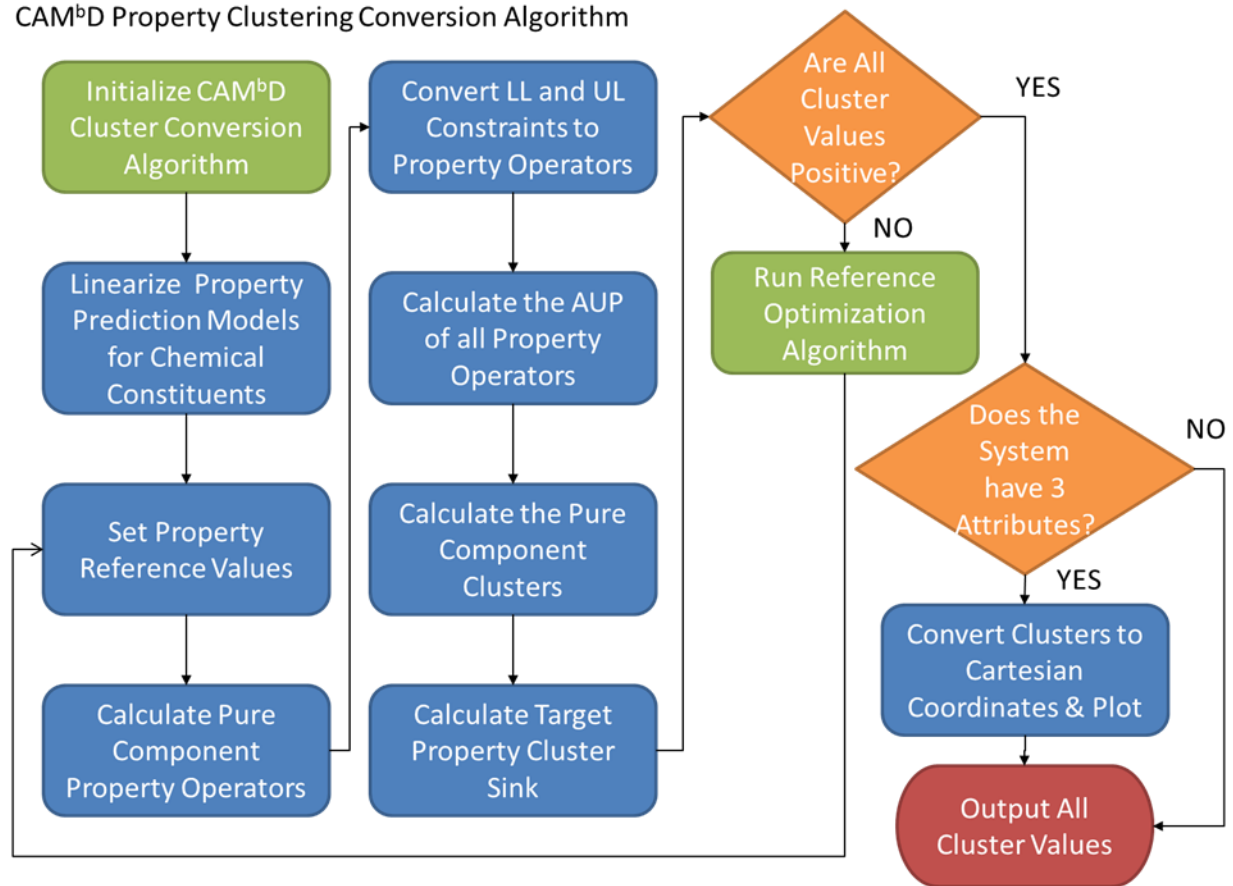


Figure 6.2: The Property Clustering Conversion Algorithm for use in Computer Aided Mixture Design (CAM^bD).

Eljack *et al.* (2007) extended the property clustering method to also include models based on group contribution methods (GCM). For this situation, Eq. 6.1 is rewritten in terms of the number of molecular groups and parameterized accordingly (Marrero and Gani 2001). If $(y_j)_g$ is the contribution to the attribute or physical-chemical property j from group g , n_g is the number of occurrences of that group in the molecule, then the molecular property operator ψ_j^M can be defined as the following:

$$\psi_j^M(y_j)_M = \sum_{g=1}^{n_g} \psi_j^M(y_j)_g \cdot n_g \quad (6.14)$$

where g is the type of group and r is the total number of first order group types. The normalized property operator may be obtained by dividing ψ_j^M with a reference value.

$$\Omega_j^M = \frac{\psi_j^M(y_j)}{\psi_j^{ref}(y_j)} \quad (6.15)$$

The Augmented Property Index AUP^M , for each molecular fragment is defined as the summation of all the j dimensionless property operators:

$$AUP^M = \sum_{j=1}^p \Omega_j^M \quad (6.16)$$

Finally, the molecular property cluster C_j^M is defined as

$$C_j^M = \frac{\Omega_j^M}{AUP^M} \quad (6.17)$$

The first order, group-specific molecular property operators described by Marerro and Gani (2001) in Table 6.1 can then be utilized in a computer aided molecular design (CAMD). Table 6.3 and Fig. 6.3 describe the property clustering conversion for systems described by group-based, attribute or physical-chemical property models.

Table 6.3. The CAMD Property Clustering Conversion Algorithm

Step	Description	Equation
1	Linearize Group Based Property Prediction Models	6.14
2	Guess Property Operator Reference Values and Calculate Group Property Operator Values	6.15
3	Convert UL and LL Feasibility Constraints to Property Operators	6.7
4	Calculate the AUP of the Group Property Operators and the Feasibility Constraint Property Operators	6.16

5	Calculate the Group Clusters	6.17
6	Calculate the Feasibility Region Sink	6.7, 6.17
7	Run Reference Optimization Algorithm for Non-Positive Clusters	6.8-6.12
8	Convert all Clusters to Cartesian Coordinates and Plot	6.23-6.26

CAMD Property Clustering Conversion Algorithm

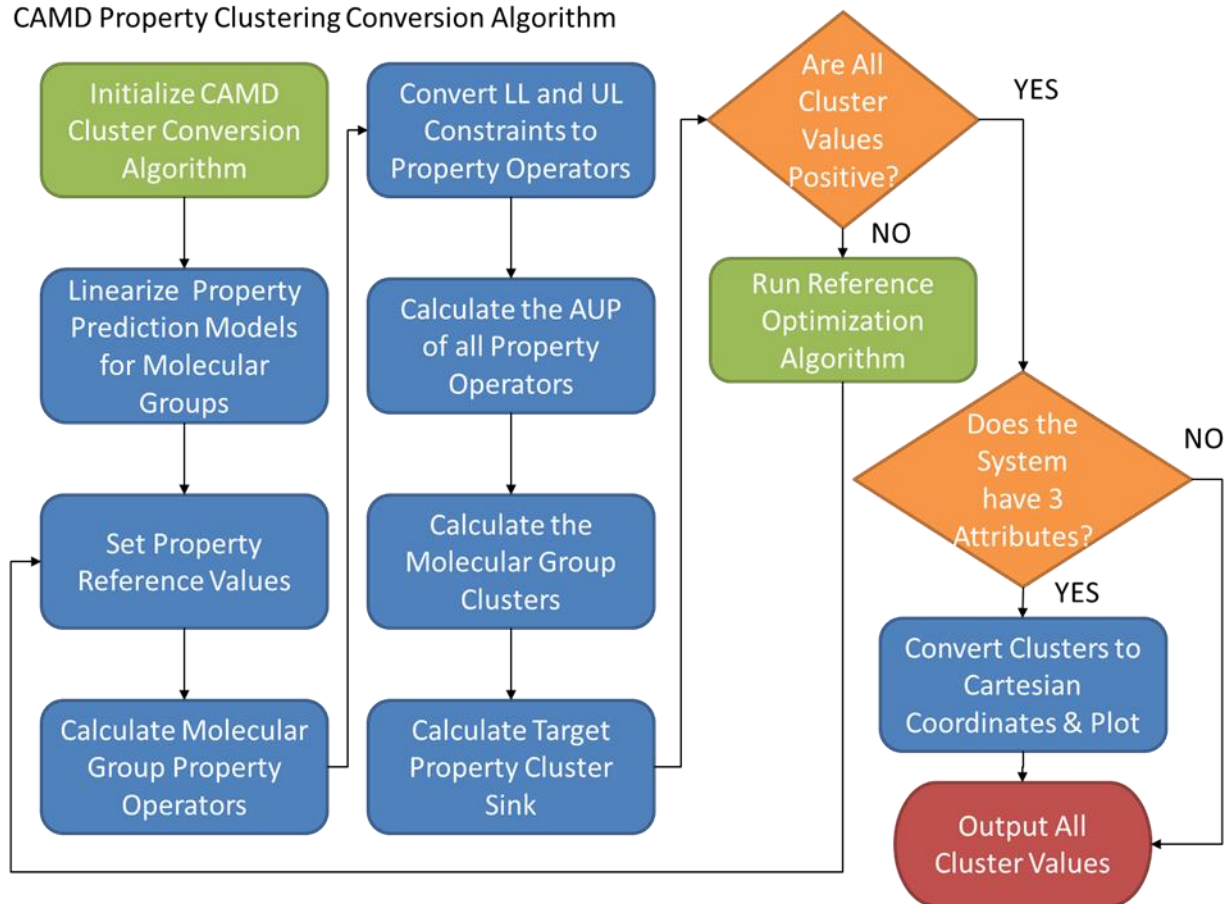


Figure 6.3: The Property Clustering Conversion Algorithm for use in Computer Aided Molecular Design (CAMD).

It should be noted that the molecular property operators defined in the above equations have limited applicability as they are formed considering only the contributions of first order molecular groups. The first order groups have limited accuracy when predicting the properties of

polyfunctional molecules with more than four carbon atoms in the main chain and cyclic molecules (Marrero and Gani, 2001). Also, first order groups cannot capture proximity effects or differentiate between isomers (Kehiaian 1983; Wu and Sandler 1989; Wu and Sandler 1991). To make this approach more robust, the molecular property operators have to be more accurate. Fortunately, higher orders of molecular groups have been introduced into GCM by Constantinou and Gani (1994) and by Marrero and Gani (2001). In their approach, the higher orders of molecular groups are formed with first order groups as their building blocks and property contributions are estimated for each group. In Chapter 8, a method utilizing first and second order group contribution within the clustering framework is presented.

6.2. Computer Aided Mixture Design (CAM^bD) with Clusters

There are two primary methods for performing chemical product design using property clustering with the reverse problem formulation: computer aided mixture design (CAM^bD) with clusters and computer aided molecular design (CAMD) with clusters. The first method is useful for finding mixtures of autonomous chemical products (e.g. molecules) while the second method is capable of finding mixtures of molecular architectures within a single chemical product (e.g. molecular groups). Shelley and El-Halwagi (2000) have shown that a property cluster is conserved through mixing which provides an opportunity to track the attribute or property response when different molecules or molecular architectures are mixed together. This intra-conserved nature of property clusters allows for the fractional calculation of any mixture in the cluster design space.

Suppose a CAM^bD is being conducted with a linear property operator expression, described by Eq. 6.1, and is used to relate properties P_1 , P_2 , and P_3 to the molecular architecture of chemical products. Using property clustering (PC), Eq. 6.1 can be rewritten as Eq. 6.18:

$$C_{jM} = \sum_{i=1}^u \delta_i \cdot C_{ji} \quad (6.18)$$

where the relative cluster arm δ_i is defined using the proof for inter-stream conservation of clusters given by Eden *et al.* (2003) and Shelley and El-Halwagi (2000). The relative cluster arm δ_i is related to changes in the augmented property index (*AUP*) during mixing, as shown in Eq. 6.19:

$$\delta_i = \frac{x_i \cdot AUP_i}{AUP_M} \quad (6.19)$$

where the AUP_M is defined in Eq. 6.20:

$$AUP_M = \sum_{i=1}^u AUP_i \cdot x_i \quad (6.20)$$

Each cluster arm δ_i can be found visually from the relative lengths of the lever arms as shown in Fig. 6.1. Alternatively, they can be calculated using Cartesian coordinates, as shown by Eden *et al.* (2003). For the three property system described by a $u=2$ component system in Fig. 6.1, the cluster arms for the point S_{mix} are as follows:

$$\delta_1 = \sqrt{\frac{(X_{CC,2} - X_{CC,M})^2 + (Y_{CC,2} - Y_{CC,M})^2}{(X_{CC,2} - X_{CC,1})^2 + (Y_{CC,2} - Y_{CC,1})^2}} \quad (6.21)$$

$$\delta_2 = \sqrt{\frac{(X_{CC,1} - X_{CC,M})^2 + (Y_{CC,1} - Y_{CC,M})^2}{(X_{CC,2} - X_{CC,1})^2 + (Y_{CC,2} - Y_{CC,1})^2}} \quad (6.22)$$

where the Cartesian coordinates (X_{CC} , Y_{CC}) are estimated from either the clusters (i.e. Eq. 6.23 and 6.24) or from the dimensionless property operators (Eq. 6.25 and Eq. 6.26).

$$X_{CC,i} = C_{1i} + 0.5 \cdot C_{2i} \quad (6.23)$$

$$Y_{CC,i} = C_{2i} \quad (6.24)$$

$$X_{CC,i} = \frac{\Omega_{1i} + 0.5 \cdot \Omega_{2i}}{AUP_i} \quad (6.25)$$

$$Y_{CC,i} = \frac{\Omega_{2i}}{AUP_i} \quad (6.26)$$

The solution to the design problem can be found by first testing each pure component against the feasibility region and then each binary mixture, ternary mixture, and so on until all combinations have been generated. A successful solution must meet three criteria (Eden *et al.* 2003; Solvason *et al.* 2009).

- Rule 1.** Cluster value of the source (or mixture of sources) must be contained within the feasibility region of the sink in the cluster domain.
- Rule 2.** The values of the Augmented Property Index (*AUP*) for the source or mixture of sources must fall within the *AUP* range of the feasibility region.
- Rule 3.** The *AUP* of the candidate must match the *AUP* of the sink at the candidate point as calculated from the sink boundary points.

In addition to these three rules, the experimenter may wish to apply supplemental constraints, such as limiting the number of pure components tested, the number components evaluated in the mixtures, or the types of mixtures considered. These constraints assume an optimization role in the design and are used to further reduce the generated list of potential candidate mixtures.

The method for testing the pure components is straightforward. As long as the pure component cluster falls within the cluster range of the sink, as defined by Eq. 6.7, and the pure

component *AUP* falls within the *AUP* range of the sink, then the *AUP* of the pure component will match the *AUP* of the sink and it will be deemed a potential candidate solution for the chemical product design. In Fig. 6.1, neither S_1 nor S_2 were found to meet the design criteria and are excluded.

Any of the previously identified candidate clusters will naturally be able to form binary mixtures with any of the other pure component cluster to meet Rule 1. The fractions of each binary mixture are determined by the location where Rules 2 and 3 are met within the feasibility sink. Since the number of pure component candidates are relatively small compared to the number of potential mixtures, most candidate mixtures will involve pure component clusters that reside outside of the feasibility region. A candidate mixture will not exist if a line segment connecting the pure component clusters does not intersect the feasibility region sink. For example, the three property system in Fig. 6.1 clearly shows that a mixture between S_1 and S_2 cannot be excluded. Rules 2 and 3 can be used to validate the existence of the candidate mixture by calculating the mass fractions of each chemical constituent. Alternatively, the Cartesian coordinates of the clusters of both the pure components and feasibility region sink can be used to test whether a line segment between the pure component cluster intersects the feasibility region. For example, a line segment between S_1 and S_2 can be written as follows:

$$A_L x + B_L y = K_L \quad (6.27)$$

$$A_L = Y_{CC,2} - Y_{CC,1} \quad (6.28)$$

$$B_L = X_{CC,1} - X_{CC,2} \quad (6.29)$$

For each Cartesian coordinate representation of the feasibility cluster point, the constant (K_L) is calculated and compared against a reference (K_L^{ref}), defined as:

$$K_L^{ref} = X_{CC,1} \cdot Y_{CC,2} - X_{CC,2} \cdot Y_{CC,1} \quad (6.30)$$

If feasibility clusters result in all $K_L < K_L^{ref}$ or all $K_L > K_L^{ref}$, then the segment does not cross the feasibility region and no mixture is possible. Otherwise, a mixture may be possible and Rules 2 and 3 will need to be validated. In the example in Fig. 6.1, a mixture between $S_1(0.3, 0.2)$ and $S_2(0.65, 0.5)$ was found to have $K^{ref} = 0.02$ and the six feasibility cluster points (Pt. 1-6) were found to have $K_L = 0.07, 0.05, 0.0, -0.03, -0.01, \text{ and } 0.04$, respectively, so the mixture cannot be excluded by Rule 1. To test Rule 2, the intersections of the line segment with the feasibility region are estimated visually or calculated using linear algebra in Cartesian coordinates as shown in Eq. 6.31, followed by estimates or calculations of the cluster arm fractions by Eq. 6.21-6.22.

$$\frac{Y_{CC,2} - Y_{CC,1}}{X_{CC,2} - X_{CC,1}} = \frac{Y_{CC,M} - Y_{CC,1}}{X_{CC,M} - X_{CC,1}} \quad (6.31)$$

The chemical constituent mass fractions of component S_1 is found by rearranging Eq. 6.19 and Eq. 6.20 to get Eq. 6.32 with the balance being S_2 .

$$x_1 = \frac{\delta_1 \cdot AUP_2}{\delta_i \cdot (AUP_2 - AUP_1) + AUP_1} \quad (6.32)$$

The AUP range of the mixture is found using Eq. 6.20 on each of the boundary intersections. If the AUP meets Rules 2 and 3, then the mixture range between the intersections is deemed a candidate solution. In situations where only one or neither of the boundary intersections meet Rules 2 and 3, a new mixture boundary point is estimated by setting the AUP_M of mixture to the AUP_F of the feasibility region at the intersection points and back calculating the cluster arms and mixture fractions of the chemical constituents. The AUP_F is chosen from the set of AUP of the cluster feasibility region points by determining which property cluster exhibits the largest change across the mixture and holding the remaining clusters constant. Alternatively, AUP_F can be

determined by mixing the feasibility region points with the mixture design to determine the point at which the mixture design satisfies the feasibility region. Once determined, the mass fraction is calculated using Eq. 6.33.

$$x_i = \frac{AUP_M - AUP_w}{AUP_i - AUP_w} \quad (6.33)$$

Negative cluster arms or mixture fractions indicate an infeasible solution and the mixture is not selected as a candidate for the chemical product design.

For ternary or larger mixtures, it is appropriate to start with the previously identified binary candidate solution clusters and add another component to the solution. Any pure component will be able to form a ternary mixture with a binary candidate solution so long as the binary candidate solution intersection with the feasibility region does not exactly lie on two of the six feasibility points, in which case a ternary mixture may not be possible. The procedure for solving for a ternary mixture is the same as for a binary mixture, with the binary cluster points forming a new feasibility region boundary for the ternary cluster mixtures, shown as $\overline{ABCDE}$ for the mixture of components S_1 , S_2 , and S_3 in Fig. 6.4.

Point D, the binary mixture point S_{12} is first calculated from the intersection of a line between $\overline{S_3D}$ and $\overline{S_1S_2}$, followed by a second binary mixture calculation between S_{12} and S_3 that intersects at Point D (or line $\overline{CD}$). Once the mixture point is found the *AUP* is calculated in the same manner as before and tested against Rules 2 and 3.

It should also be noted that although binary mixtures of $S_2 - S_3$, $S_3 - S_4$, and $S_2 - S_4$ were excluded as candidate solutions in the example, a ternary mixture of S_2 , S_3 , and S_4 would completely encompass the feasibility region and would meet the Rule 1, making it a potential candidate solution. To determine if the ternary mixture encompasses the feasibility region, the maximum and minimum cluster values of the points (e.g. Eq. 6.7) that make up the feasibility boundary are calculated from Eq. 6.34 and 6.35.

$$C_j^U = \max_i (C_{ji}) \quad (6.34)$$

$$C_j^L = \min_i (C_{ji}) \quad (6.35)$$

In a similar fashion, a mixture cluster range is determined from the maximum and minimum cluster values of the pure components in the ternary mixture. If the maximum cluster of the ternary mixture is not larger than the maximum cluster of the feasibility region or the minimum cluster of the mixture is not smaller than the minimum cluster of the feasibility region, then the mixture does not overlap the feasibility region and is discarded. If the mixture does overlap the feasibility region, then the solution to this type of problem is identical to that of the first condition, only that the entirety of feasibility region boundary points will need to be calculated using the ternary mixture procedure. As more chemical constituents are used in the design of the chemical product (i.e. $u \geq 3$) complete overlaps, like $\overline{S_2S_3S_4}$ overlapping the feasibility region in Fig. 6.4, become more common. Although not shown here, chemical product designs with more than 3 attributes or properties can also be solved using property clustering. This situation

requires the use of an algebraic approach developed by Qin *et al.* (2004) and Eljack *et al.* (2007), but is beyond the scope of this dissertation.

In conclusion, Table 6.4. and Fig. 6.5-6.8 present a summary of the computer aided mixture design (CAM^bD) method using property clustering. To begin using the algorithm, the researcher must set the maximum number μ^{max} of dissimilar chemical species μ utilized in a mixture. This constraint helps to determine if the mixtures will be pure components, binary, ternary, quarternary, etc. A similar constraint is used in the computer aided molecular design (CAMD) and is discussed in more detail in Section 6.3. Next, pure component solutions are found by testing the cluster outputs of the CAM^bD conversion algorithm against Rules 1-3. Any pure component candidates are saved into a candidate solution database. Next, binary mixtures with at least one pure component candidate are calculated and tested against Rules 1-3, with candidates being uploaded into the candidate solution database. This step is followed by calculating and testing binary mixtures with no pure component candidates against Rules 1-3 and outputting the binary mixtures to the candidate solution database. The procedure is repeated for ternary and larger mixtures, beginning with mixtures that contain binary mixture candidates, then those without, and outputting all candidate mixture into the database.

Table 6.4. The CAM^bD Candidate Generation Algorithm using Clusters

Step	Description	Equation
1	Set μ^{max} Maximum Dissimilar Species in Mixture	-
2	Discard Pure Component Clusters (from CAM ^b D Conversion Algorithm) that Fail Rule 1	6.6, 6.7
3	Discard Pure Component Clusters that Fail Rule 2, Output Pure Component Candidate Solutions	6.5, 6.7
4	For Binary Mixtures with Pure Component Candidates, Calculate the Clusters at the Intersections of the Sink and the Mixtures	6.31, 6.23, 6.24

5	Calculate the Binary Cluster Arms	6.21, 6.22
6	Calculate the Fractions of Each Pure Component	6.32
7	Calculate the AUP of Mixtures, Output Mixtures that Meet Rules 2 and 3 to Candidate Library	6.20
8	For Mixtures that Fail Rule 2 or 3, set the AUP_M at the UL or LL of the Sink and Recalculate the Component Fractions, Discard Binary Mixtures with Component Fractions < 0 or > 1	6.20
9	Recalculate Cluster Arms of Candidate Mixture	6.21, 6.22
10	Recalculate Candidate Cluster and Output Binary Candidate Mixtures	6.18
11	For Binary Mixtures with no Pure Component Candidates, Calculate the Reference Constant for each Mixture	6.23, 6.24, 6.30
12	Calculate the Sink Constants for each Mixture, Discard Mixtures where all Sink Constants are $>$ or $<$ Reference Constant	6.23, 6.24, 6.27-6.29
13	Calculate the Clusters at the Intersections of the Sink and the Mixture and Repeat Steps 5-9	6.31, 6.23, 6.24
14	For $\mu \geq 3$ Ternary+ Mixtures with $\mu - 1$ Candidates, Calculate the $\mu - 1$ Equivalent Clusters	6.31, 6.23, 6.24
15	Calculate the Clusters at the Intersections of the Sink and the $\mu \geq 3$ Mixture and Repeat Steps 5-9, Output $\mu \geq 3$ Mixtures that Meet Rules 2 and 3	6.31, 6.23, 6.24
16	For Mixtures that Fail Rule 2 or 3, set the AUP_M at the UL or LL of the Sink and Recalculate the Component Fractions, Discard Mixtures with Component Fractions less than 0 or greater than 1	6.20
17	Recalculate Cluster Arms of Candidate Mixture	6.21, 6.22
18	Recalculate Candidate Cluster and Output $\mu \geq 3$ Candidate Mixtures to Library	6.18
19	For $\mu \geq 3$ Ternary+ Mixtures with no $\mu - 1$ Candidates, Calculate the Cluster Range of the Pure Components	6.34, 6.35, 6.6
20	Calculate the Cluster Range of the Feasibility Sink, Discard Mixtures whose Pure Component Cluster Range is $>$ Sink Cluster Range	6.34, 6.35, 6.6, 6.7
21	Repeat Steps 14-20 until μ^{max} is reached	-

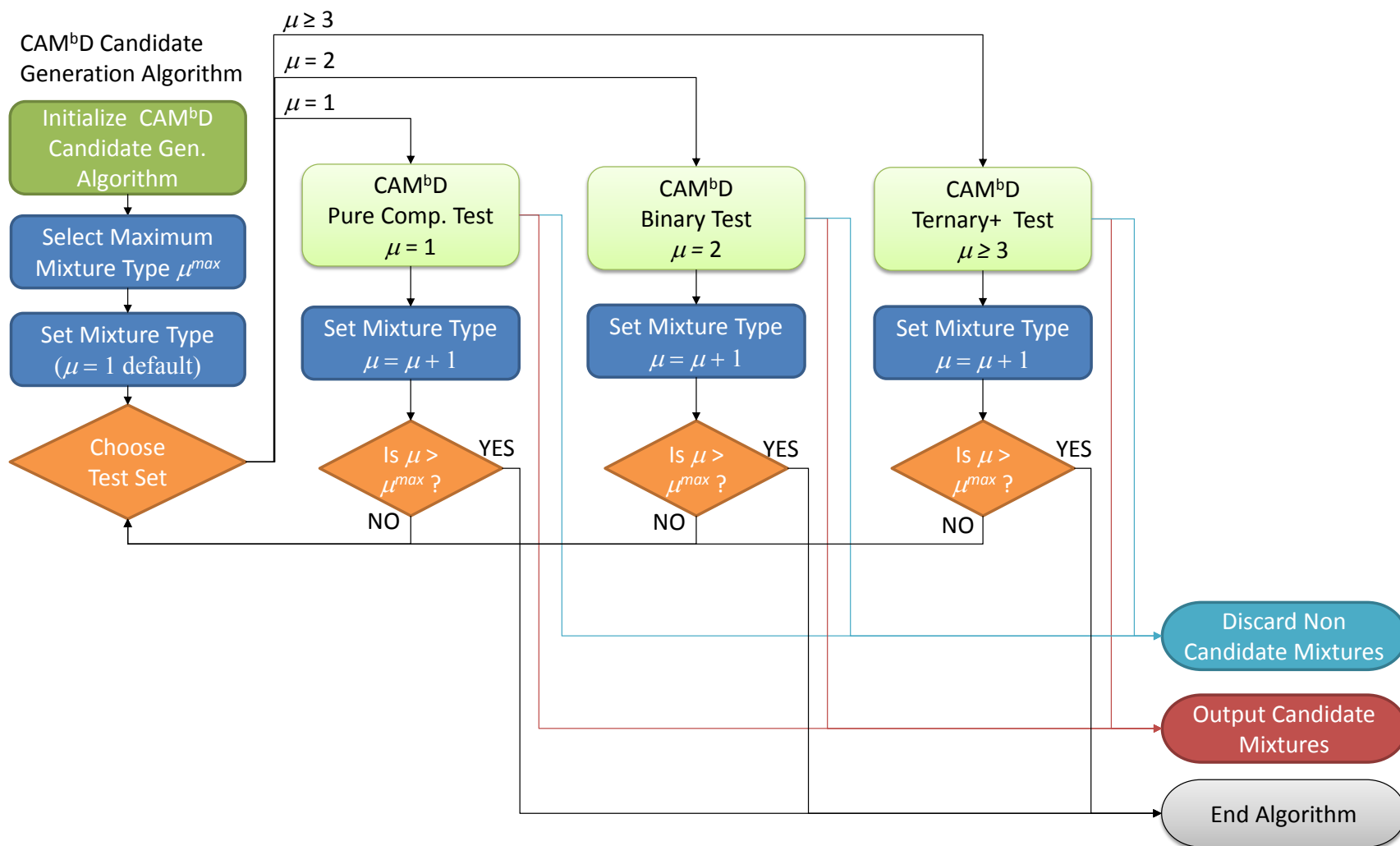


Fig. 6.5 Computer Aided Mixture Design (CAM^bD) Candidate Generation Algorithm using Property Clustering.

Each of the tests shown in Fig. 6.5 can be defined separately. Fig. 6.6 describes the CAM^bD cluster test on pure components. Fig. 6.7 describes the CAM^bD cluster test on binary mixtures, starting with any pure component candidates identified in Fig. 6.6. Fig. 6.8 describes the CAM^bD cluster test on ternary and larger mixtures, starting with any $\mu-1$ candidate mixtures identified.

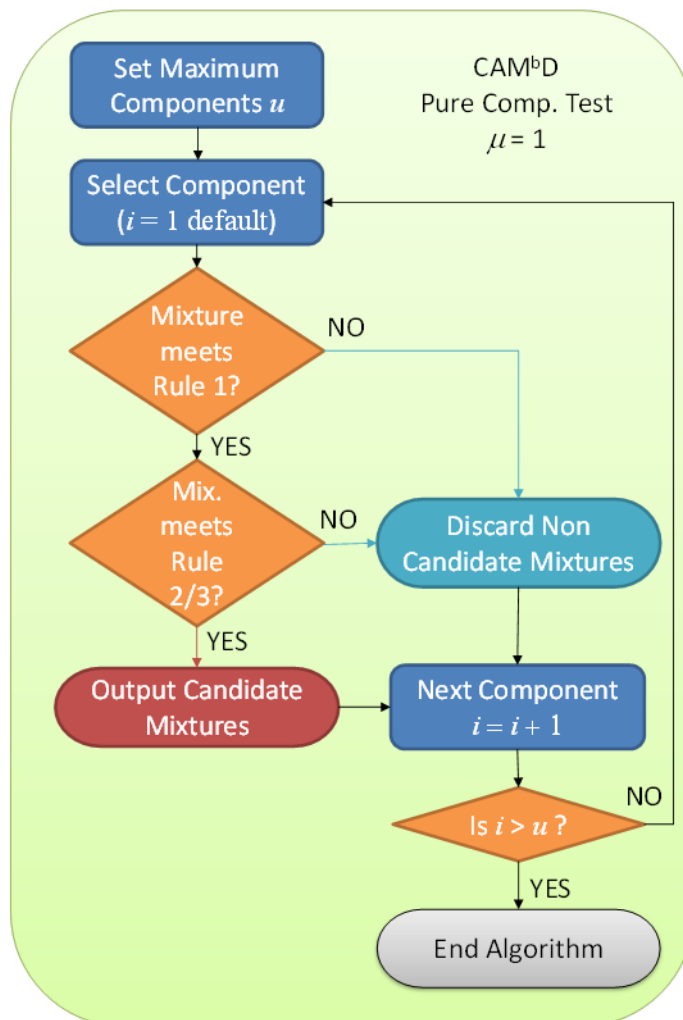


Figure 6.6: CAM^bD with Clusters test on Pure Components

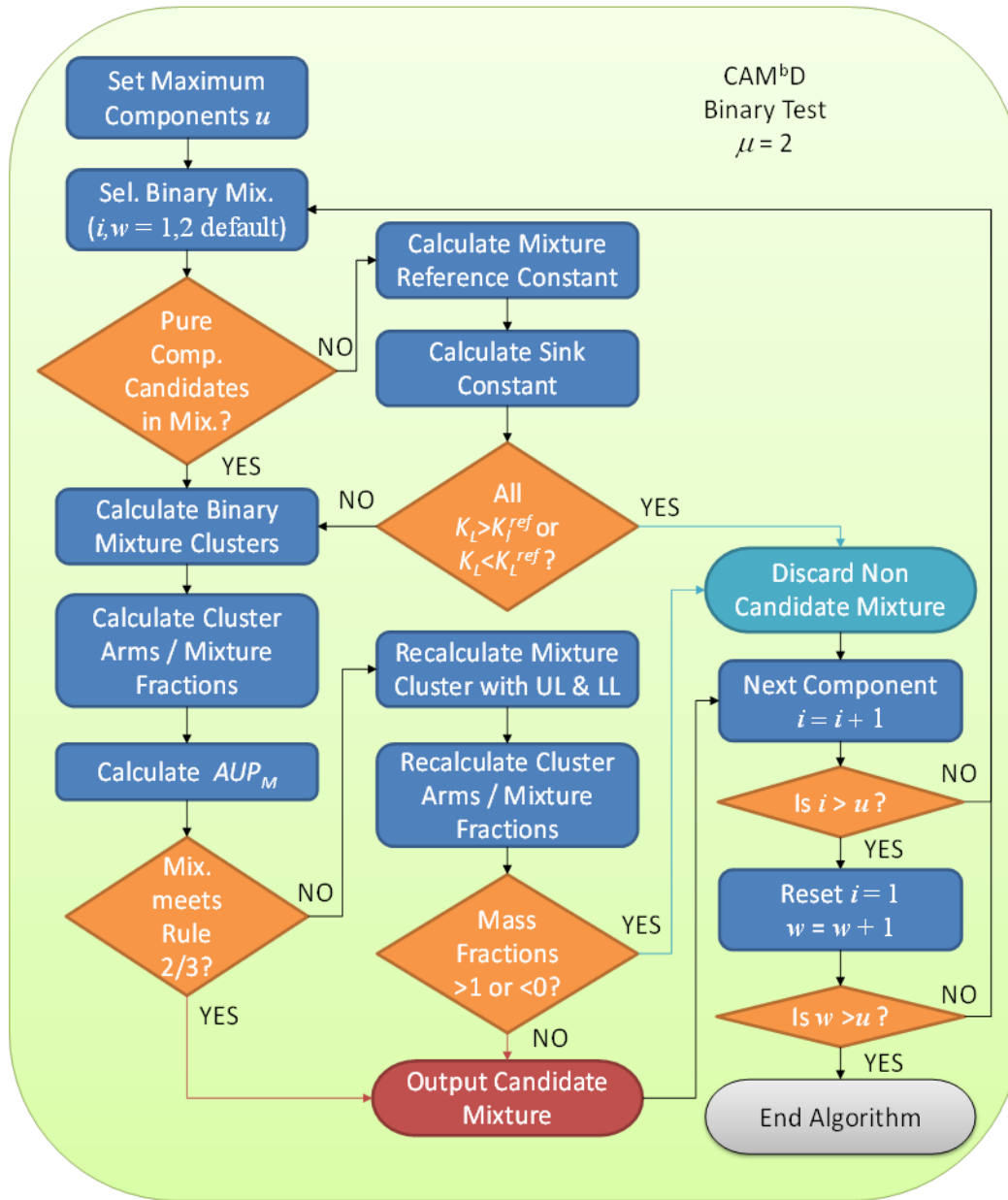


Figure 6.7: CAM^bD with Clusters test on Binary Mixtures

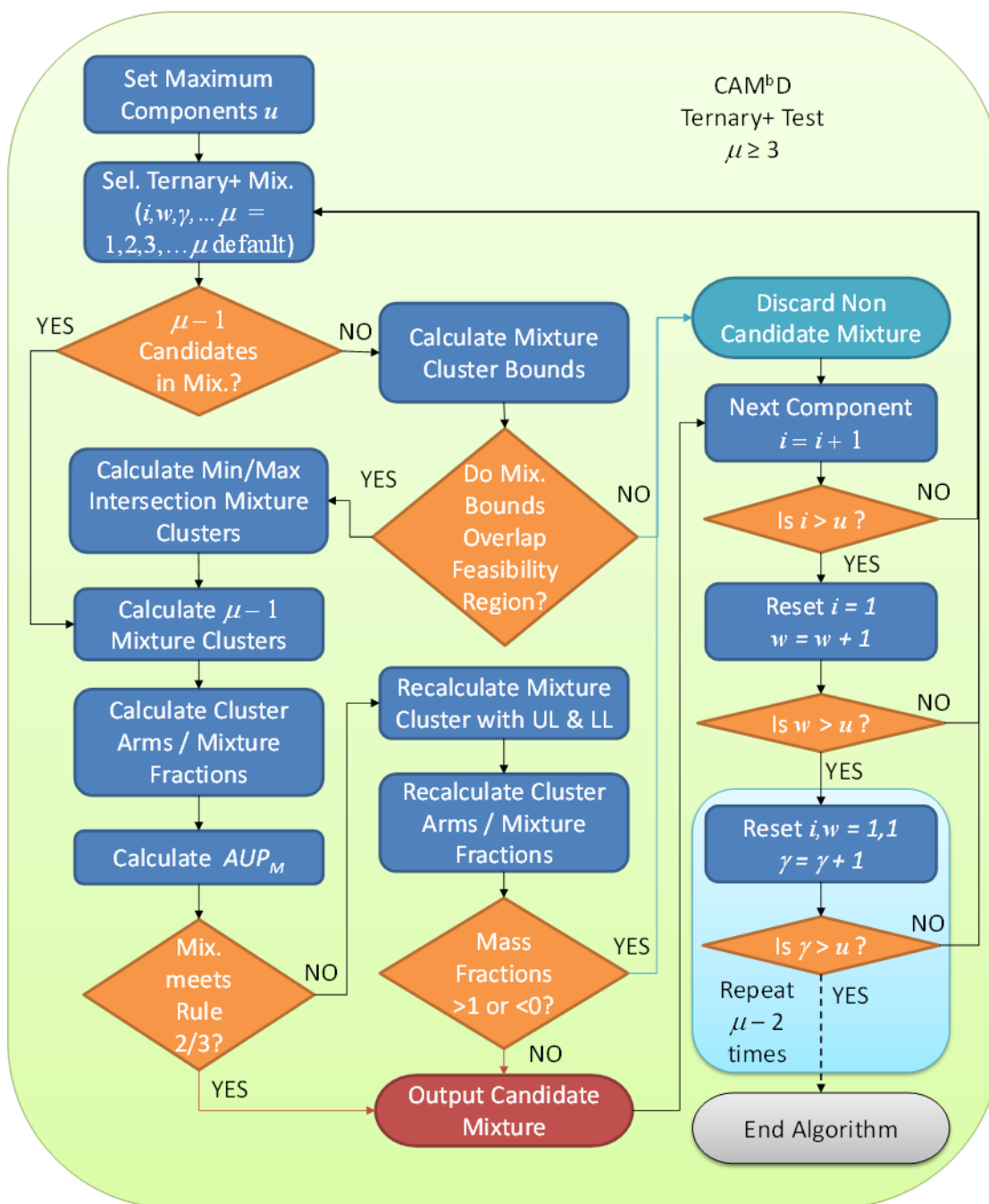


Figure 6.8: CAM^bD with Clusters test on Ternary and Larger Mixtures

6.3. Computer Aided Molecular Design (CAMD) with Clusters

The second method for performing chemical product design using property clustering is computer aided molecular design (CAMD) with clusters. While CAM^bD is useful for finding

mixtures of autonomous chemical products (e.g. molecules) that meet product specifications, CAMD is capable of finding mixtures of molecular architectures (e.g. molecular groups) that make single chemical products (e.g. molecules) that meet product specifications. Eden *et al.* (2003) and Eljack *et al.* (2007) have shown that first order groups can be easily expressed linearly in Eq. 6.14, making them ideal candidates for use in CAMD with clusters.

In order to perform CAMD with clusters, the property clusters must be conserved during mixing. Eljack *et al.* (2007) has shown that first order molecular property clusters possess both intra-stream and inter-stream conservation which allows for the tracking of attribute or property responses during molecule generation. The group contribution parameters are regressed from the CAPEC database (Marrero and Gani, 2001) on a group basis and represented as clusters according to Eq. 6.15-6.17. The intra-stream molecular conservation results in Eq. 6.36.

$$\sum_j^p C_j^M = \frac{\sum_j^p \Omega_j^M}{AUP^M} = \frac{AUP^M}{AUP^M} = 1 \quad (6.36)$$

Inter-molecular conservation requires that the individual group clusters C_j^M be conserved during mixing. The property operator expression of Eq. 6.14 can be non-dimensionalized by Eq. 6.15, resulting in Eq. 6.37.

$$\Omega_{ji}^M = \sum_{g_1=1}^{n_g} n_{g_1} \Omega_{jg_1}^M \quad (6.37)$$

Rewriting the molecular property operator expression in terms of clusters is done by inserting Eq. 6.37 into Eq. 6.16 and rearranging to get Eq. 6.38-6.40.

$$C_{ji}^M = \frac{\sum_{g_1}^F n_{g_1} \cdot \Omega_{jg_1}^M}{AUP_i^M} = \frac{\sum_{g_1}^F n_{g_1} \cdot C_{jg_1}^M \cdot AUP_{g_1}^M}{AUP_i^M} \quad (6.38)$$

$$\delta_{g_1} = \frac{n_{g_1} \cdot AUP_{g_1}^M}{AUP_i^M} \quad (6.39)$$

$$C_{ji}^M = \sum_{g_1=1}^F \delta_{g_1} \cdot C_{jg_1}^M \quad (6.40)$$

The group based property model shown in Eq. 6.40 is parameterized as first order groups by Eljack *et. al.* (2007) and demonstrates how molecular groups or fragments can be added together analogous to inter-stream conservation. Higher order group contribution can increase the accuracy of the predicted properties and will be discussed in more detail in Chapter 8.

The ternary diagram used for mixture design as a visualization tool for source-sink mapping can again be used to highlight the molecular design method. In the cluster formulation for CAMD, mixing of two sources is a straight line, i.e. the mixing operation can be optimized using lever-arm analysis. Analogously, combining or “mixing” two molecular fragments in the molecular cluster domain follows a straight line as shown by a mixture of G1 and G2 in Fig. 6.9. This result can be used to design molecules that meet Rule 1 (e.g. G1-G2-G3 and G1-G2-G3-G4), such that the length of each cluster arm δ_g can be calculated from Eq. 6.38 and the number of that group, n_g , can be found. However, before Rules 2 and 3 can be used to narrow the candidate set of molecules in the design, an additional rule unique to molecular design needs to be included. In particular, it is important to ensure that the designed molecule is electronically complete, meaning that its Free Bond Number (*FBN*) shown in Eq. 4.9 must be 0. In terms of first order molecular groups, the expression becomes Eq. 6.41.

$$FBN = \sum_{g_1=1}^F n_{g_1} FBN_{g_1} - 2 \cdot \left(\sum_{g_1=1}^F n_{g_1} - 1 \right) - 2 \cdot N_{RG} \quad (6.41)$$

where the FBN is the free molecular bond number of the formulation, n_{g1} is the number of occurrences of group g_1 , FBN_{g1} is the unique free bond number associated with group g_1 , N_{RG} is the number of rings in the formulation, and F is the total number of first order groups. Forcing Eq. 6.40 to be 0 is stated as Rule 4.

Rule 4. The Free Bond Number (FBN) of the candidate molecule must be zero.

It should be noted that the location of the final formulation is independent of the order of group addition, making it ideal for use in computational algorithms.

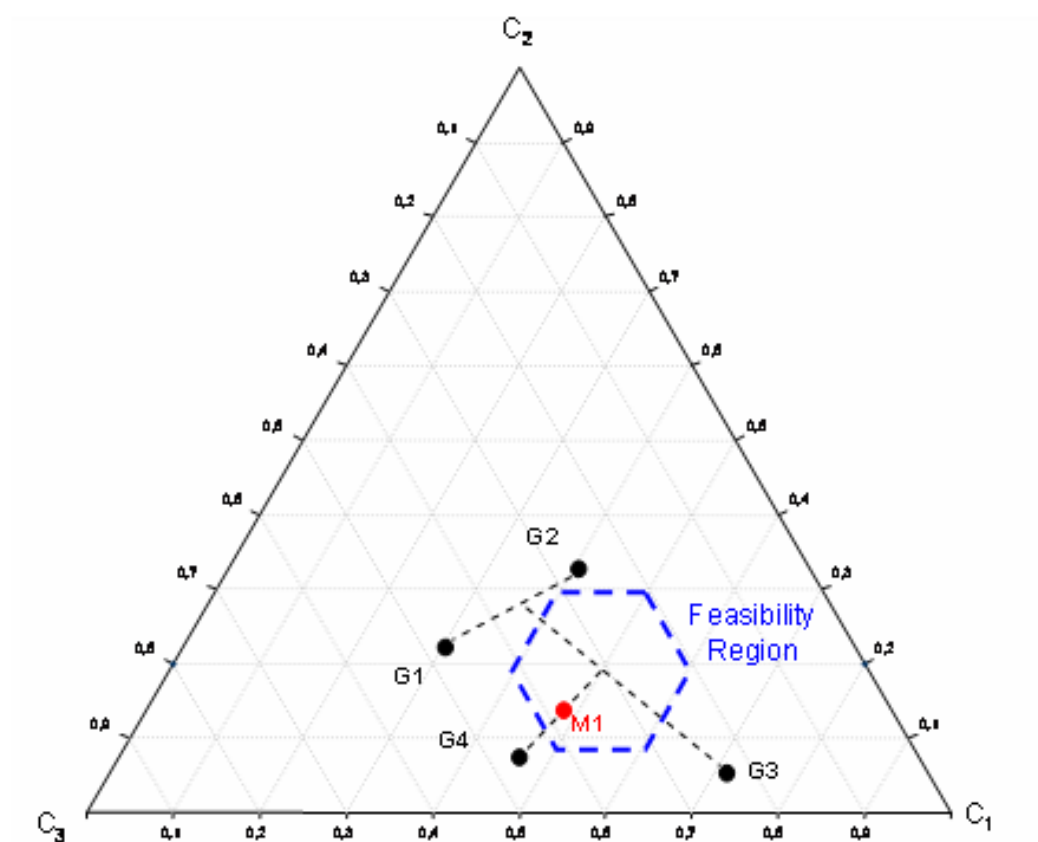


Figure 6.9: Group Addition on a Ternary Cluster Diagram using Four Dissimilar Molecular Groups to Build a Molecule (Eljack *et al.* 2007).

As shown in Fig. 6.10, the method proceeds in a similar manner to that of CAM^bD, but with groups instead of components as the combinatorial building blocks. The maximum number

of dissimilar groups μ_g^{max} is first set and then progressively larger combinations of dissimilar groups are added until the maximum is reached. As shown in Fig. 6.11 for “pure group” mixtures (e.g. $\mu_g = 1$), the pure group property contributions are estimated from Eq. 6.38 for each g_1 group type. These clusters are then tested against the points making up the feasibility bounds of a target feasibility region (FR) defined using the clustered versions of the points in Eq. 6.7. All “pure group” property clusters (e.g. Eq. 6.17) that meet Rule 1 are then passed on to test against Rule 2. For group based design, Rule 2 can be created by setting Eq. 6.39 equal to 1 for a “pure group” and resulting in Eq. 6.42

$$AUP^L \leq n_{g_1} AUP_{g_1} \leq AUP^U \quad (6.42)$$

Hence, unlike mixture designs using pure components, it is possible adjust the number of groups in the molecular structure in order to meet the constraint of Eq. 6.42. This result can be used to calculate the maximum and minimum number of groups of each type that can be utilized in a $\mu_g = 1$ design.

$$n_{g_1}^U \leq \text{int} \left(\frac{AUP^U}{AUP_{g_1}} \right) \quad (6.43)$$

$$n_{g_1}^L \geq \text{int} \left(\frac{AUP^L}{AUP_{g_1}} \right) \quad (6.44)$$

$$n_{g_1}^L \leq n_{g_1} \leq n_{g_1}^U \quad (6.45)$$

By default, any pure group mixtures that meet Rule 2, will also meet Rule 3. Any mixtures that fail Rule 2 are discarded. Finally the number of groups n_{g_1} are indexed from $n_{g_1}^L$ to $n_{g_1}^U$ and tested against Rule 4 using Eq. 6.41. Any solution that fails Rule 4 is summarily discarded. Molecules that meet Rule 4 are outputted as candidate solutions.

As shown in Fig. 6.12 for binary mixtures of dissimilar types of groups (i.e. $v_1 \neq g_1$) in non ring compounds (i.e. $\mu_g = 2$, $N_{RG} = 0$), the CAMD procedure follows a similar path as that for CAM^bD, beginning first by testing mixtures of dissimilar groups that contain one pure group candidate solution (e.g. a mixture of M1 and G4). Since the number of groups is discrete, it is generally easier to proceed in a forward manner by assuming Rule 1 is met and varying the combinations of the number of groups, testing each mixture sequentially against Rule 2 - Rule 4, and then validating that Rule 1 is indeed satisfied. The procedure starts by building candidate molecules by indexing the number of groups n_{g1} from a minimum of 1 to a maximum defined by Eq. 6.43 for each group g_1 . Then, the number of the dissimilar group n_{v1} is indexed and the first step is repeated. The procedure continues until the number of a dissimilar group reaches a maximum as defined by Eq. 6.46.

$$n_{v_1}^U \leq \text{int}\left(\frac{AUP^U}{AUP_{v_1}}\right) \quad (6.46)$$

The identified combinations are then tested against rule Rule 2 using Eq. 6.42. By definition, any molecules that meet Rule 2 will automatically meet Rule 3, so they are immediately passed forward to be tested for molecular completeness using Rule 4. Combinations that fail Rule 2/3 or Rule 4 are discarded while combinations that satisfy Rule 4 are passed to the final validation step. The validation of Rule 1 is performed by testing the clusters of the mixture solutions against the FR clusters. Mixture combinations that fall inside the FR are outputted as candidate solutions while those that fall outside the FR are discarded.

Since the number of pure group candidates are relatively small compared to the number of potential molecules, most candidate mixtures will involve pure group clusters that reside outside of the feasibility region. As with the CAM^bD with clusters algorithm, a candidate

mixture will not exist if a line segment connecting the pure clusters does not intersect the feasibility region sink (FR). Using the Cartesian coordinates of the clusters of the pure groups and feasibility region, the presence of a line segment between pure groups that intersects the FR can be tested using Eq. 6.27-6.30. If the intersection is not confirmed, like between pure groups G1 and G2, then all molecular combinations of only those dissimilar groups are discarded. If an intersection is confirmed, like theroretically between pure groups G1 and G3 in Fig. 6.9, the binary candidate generation steps are repeated using Eq. 6.47.

$$AUP^L \leq \sum_{g_1=1}^F n_{g_1} AUP_{g_1} \leq AUP^U \quad (6.47)$$

The implicit structure of this expression prevents universally setting constrained upper and lower limits on the number of groups formed beyond the limits proposed in Eq. 6.45. As a result the number of groups in the potential candidate mixtures that meet rule 1 are set using Eq. 6.48:

$$1 \leq n_{g_1} \leq n_{g_1}^U \quad (6.48)$$

Each of the candidate solutions is then tested against Rule 2 and Rule 3 using Eq. 6.47. As with the pure group solutions, binary solutions that meet Rule 2 and 3 are then passed forward to the structure validation step, followed by a validation step using Rule 1. Binary candidate and non-candidate mixtures are outputted accordingly. Any repeated solutions are also discarded (e.g. a binary mixture of G1-G2 can also be represented as G2-G1).

For ternary and sequentially larger combinations of dissimilar groups (e.g. ternary+), it is appropriate to start with the previously identified $\mu_g - 1$ candidate solution clusters and add another pure group to the solution. All ternary+ combinations of the set of n_{g1} within the bounds of Eq. 6.48 are tested against Eq. 6.47. Ternary solutions that meet Rule 2 and 3 are then tested for structural completeness using Rule 4 and vailidated using Rule 1.

Ternary+ solutions that do not contain any $\mu_g - 1$ candidate solutions can potentially encompass the feasibility region. The maximum and minimum cluster values of these pure component solutions are tested against the maximum and minimum cluster values of the FR, shown in Eq. 6.34 and 6.35. If the solution does overlap the feasibility region, then the procedure to solve this type of problem is identical to that of the binary procedure. Solutions that fail to overlap the feasibility region are discarded as non-candidate solutions.

The procedure is repeated until a maximum number of dissimilar groups μ_g^{max} is reached. A summary of the method is provided in Table 6.5 and Fig. 6.10-6.14.

Table 6.5: The CAMD Candidate Generation Algorithm using Clusters

Step	Description	Equation
1	Set Maximum Dissimilar Groups in the Mixture	-
2	Discard Pure Group Clusters (from CAMD Conversion Algorithm) that Fail Rule 1	6.17, 6.7
3	Calculate No. Pure Groups	6.45
4	Calculate Pure Group Solutions, Discard Pure Group Solutions that Fail Rule 2 and Rule 3	6.7, 6.16, 6.42
5	Discard Pure Group Clusters that Fail Rule 4, Output Pure Group Candidate Molecules	6.41
6	For Binary Solutions with Pure Group Candidates, Calculate the No. Binary Groups	6.46, 6.48
7	Calculate Solution AUP_M^M , Discard Solutions that Fail Rule 2	6.39, 6.47
8	Calculate the FBN, Discard Solutions that Fail Rule 4	6.41
9	Calculate Solution Clusters, Discard Solutions that Fail Rule 1, and Output Candidate Solution Clusters	6.7, 6.17
10	For Binary Solutions with no Pure Group Candidates, Calculate the Reference Constant for each Mixture	6.20
11	Calculate the Sink Constants for each Mixture, Discard Mixtures where all Sink Constants are $>$ or $<$ Reference Constant, Repeat Steps 6-9	6.23-6.24, 6.27-6.29

12	For Ternary and Larger Solutions ($\mu_g \geq 3$) containing $\mu_g - 1$ Candidates, Calculate the No. Groups, Repeat Steps 7-9	6.46, 6.48
13	For $\mu \geq 3$ Ternary+ Mixtures with no $\mu - 1$ Candidates, Calculate the Cluster Range of the Pure Components	6.34, 6.35, 6.6
14	Calculate the Cluster Range of the Feasibility Sink, Discard Mixtures whose Pure Component Cluster Range is $>$ Sink Cluster Range, Repeat Step 12	6.34, 6.35, 6.6, 6.7
15	Repeat Step 12-14 until μ_g^{max} is reached	-

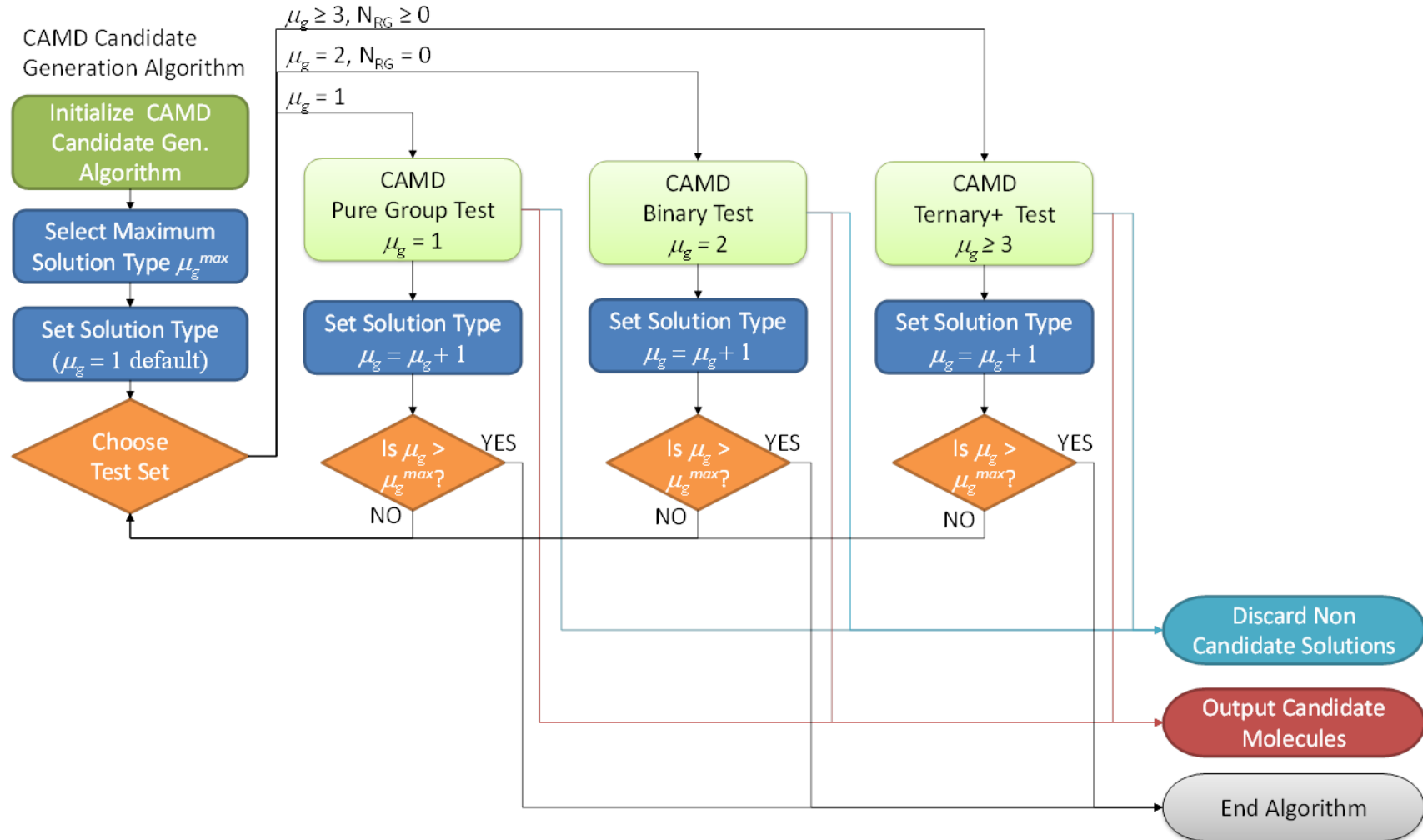


Figure 6.10: A Flowchart of the CAMD Candidate Generation Algorithm using Clusters

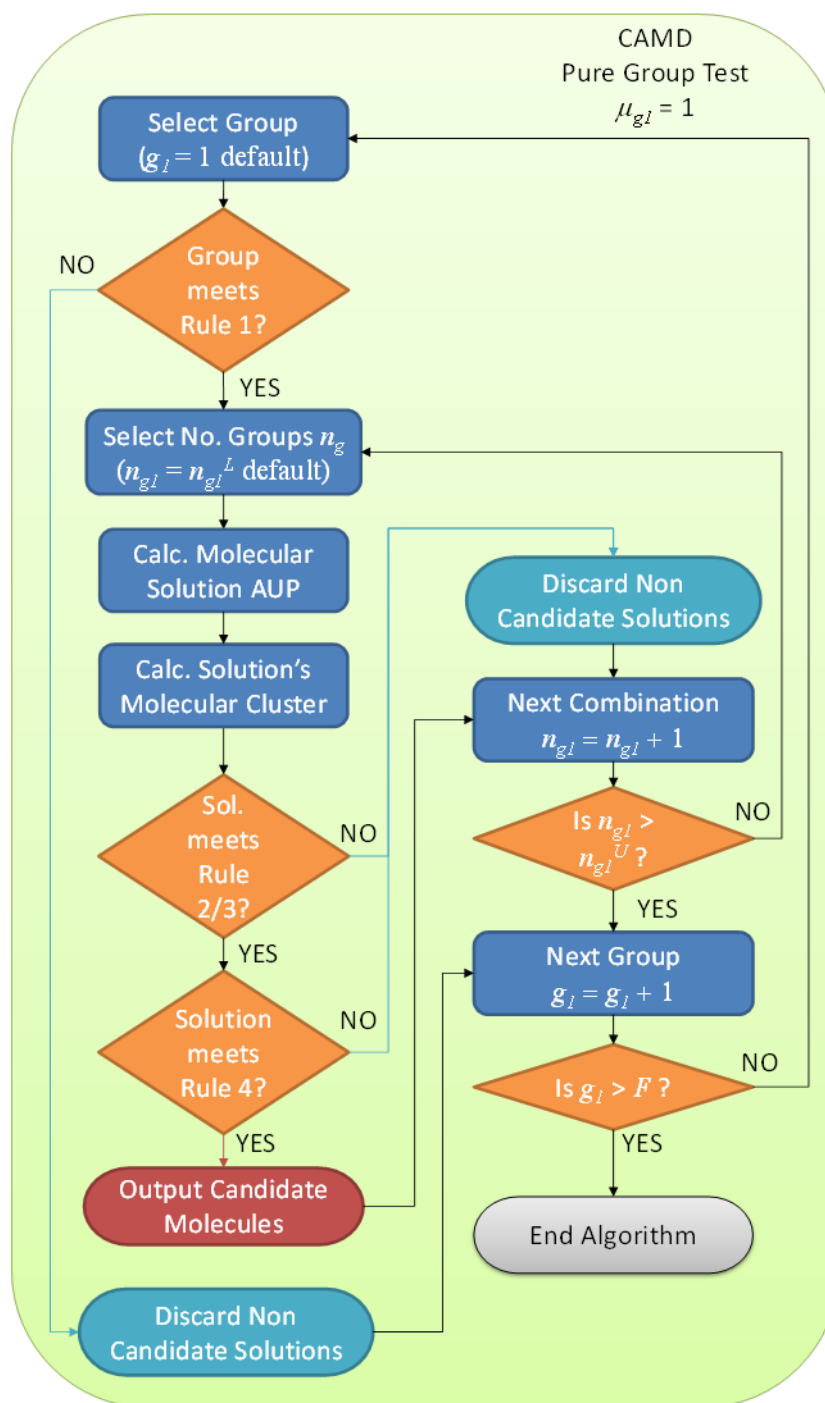


Figure 6.11: Generation of Pure Group Candidate Molecules

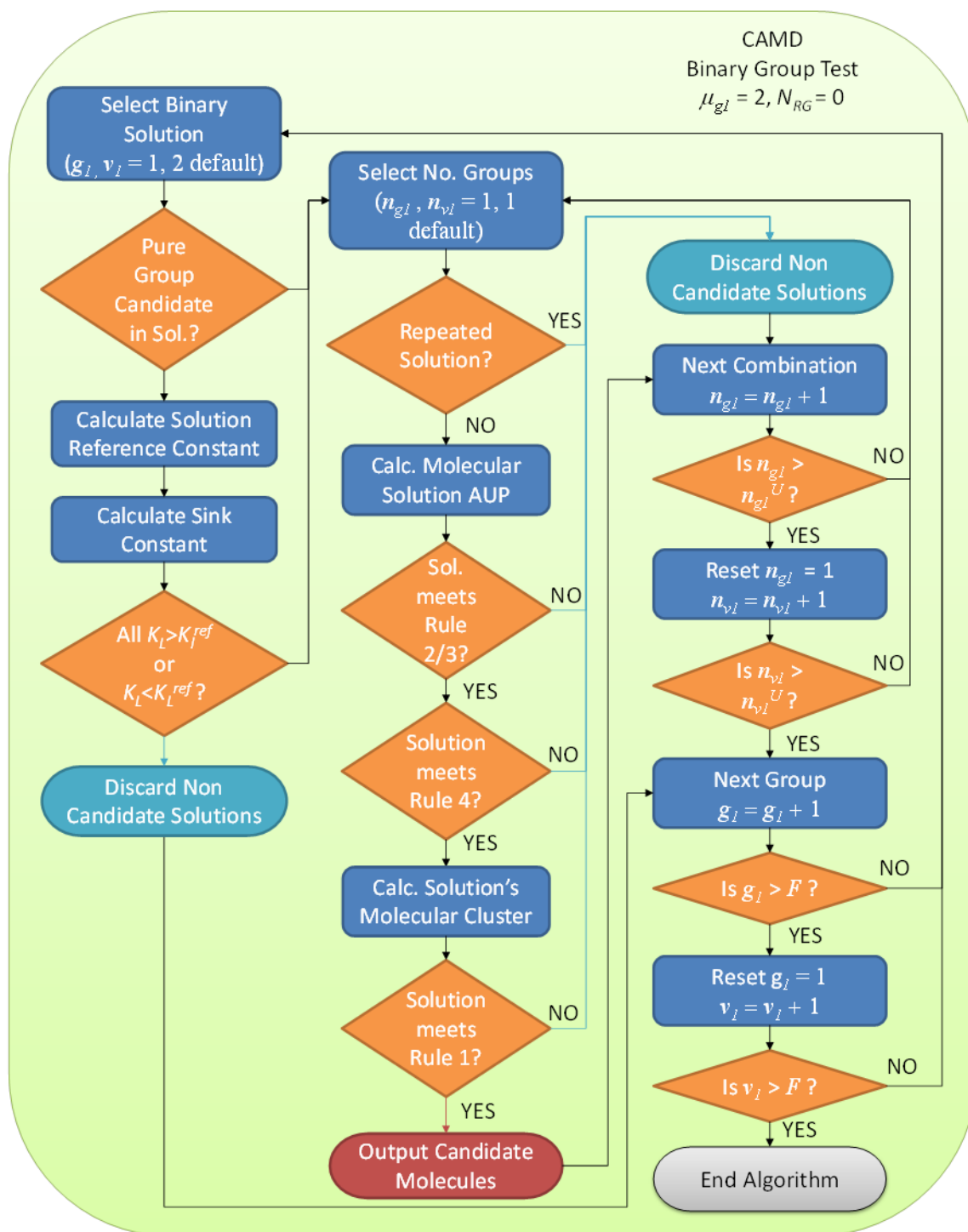


Figure 6.12: Generation of Binary Group Candidate Molecules

6.4. Summary

In conclusion, the complexity of chemical product design problems can be reduced by using the reverse problem formulation and the property clustering algorithm in a number of ways. First, by solving the design problems in the lower dimensional property domain, an unlimited number of molecular architecture solutions can be evaluated at minimal computational cost. Second, the multitude of property response plots normally associated with chemical product design can be reduced to $p - 2$ simplex diagrams to aide visualization of the design problem. Property clustering also consolidates the numerous effects of components or groups on the response into distinguishable values which may be used to surmise which of the components or groups has the largest effect and is thus most important. Both of these property clustering attributes can be used to guide chemical product designs toward feasible solutions more quickly than conventional techniques.

Since the objective of the research presented in this dissertation is to successfully link the multiple length scales in a reverse problem formulation, use of the CAM^bD and CAMD cluster algorithms to consolidate the large amount of information and improve computational efficiency will be essential. In Chapter 5, it was noted that such a design problem will invariably utilize data due to the lack of explicit models that describe consumer preference behavior. Based on this observation, Chapter 7 explores mapping experimentally derived models and data into the property clustering structure and how property clustering can be used to improve the design capability of such systems, resulting in a new method called attribute based computer aided mixture design (aCAM^bD). Although the technique developed significantly improves efficiency, it is subject to choosing only the best candidate mixtures from a predetermined training set and incapable of looking at new molecular structures. In Chapter 8, another new method is

developed to address this shortcoming by mapping the experimental data into a domain that is described by known combinatorial property prediction methods, opening the door to true computational molecular design. This new method is known as attribute based computer aided molecular design (aCAMD). When good mapping functions cannot be derived or the computational methods are too restrictive, Chapter 9 presents a work around that maps the experimental data into a sub-domain capable of handling both experimental design, computer aided molecular design, and structure design techniques through the use of characterization data. The additional benefit of this technique is that all length-scales important to the chemical product can be mapped down to the sub-domain, representing the ultimate in combinatorial efficiency and is the primary result of this dissertation. This technique is noted as characterization based computer aided molecular design (cCAMD).

Chapter 7

Attribute-Computer Aided Mixture Design (aCAM^bD)

The terms product synthesis and design designate problems involving identification and selection of compounds or mixtures that are capable of performing certain tasks or possess certain physical properties. 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. However, the performance requirements for the desired component(s) are usually dictated by the process and thus the identification of the desired component properties should be driven by the desired process performance. Whereas numerous contributions have been made in the areas of molecular synthesis and computer aided molecular design (CAMD) by Harper *et al.* (1999), Harper and Gani (2000), Marcoulaki and Kokossis (1998), and Eljack *et al.* (2007) among others, little focus has been on utilizing experimental design techniques when attribute or physical-chemical property prediction tools are insufficient. This chapter highlights a data driven Computer Aided Mixture Design (CAM^bD) method, known as attribute CAM^bD or aCAM^bD, that is capable of handling experimental data and regression models within the property clustering framework. In Section 7.1, the constraints on the use of attribute and property data within the property clustering framework is discussed and the two most common regression models, canonical and polynomial, are investigated as prediction tools. In Section 7.2 the attribute clustering technique is reviewed and new rules are added to handle negative regressors. Section 7.3 describes the aCAM^bD method using property clusters and

Section 7.4 highlights the method in a case study on the polymer blend of spun yarn. Section 7.5 concludes the chapter with a summary.

7.1. Integrating Attribute Data and Models with Property Clustering

Early in experimental mixture design, Scheffe (1958, 1963) and Cox (1971) developed techniques to obtain property models while minimizing experimental runs or design points utilizing simplex diagrams of the chemical constituent design space. However, finding the solution in the chemical constituent design space is prone to combinatorial explosion. Visualizing and solving the problem in the property space avoids this problem while also offering insights into the effectiveness of the design. Thus, the overall objective of this contribution is to integrate the property clustering framework with existing mixture design techniques used to develop data driven property models. The two most common property-mixture designs, Scheffe canonical models and Cox polynomial models, are evaluated. The results of the exercise will be used to develop additional techniques for utilization of PCR and PLS models under combinatorial explosion.

As discussed in Chapter 3, the most common approach to efficiently develop data driven models is to use Design of Experiments (DOE). In DOE, a model is first postulated to represent a potential property response surface. Next, experimental design points are placed in areas where observations can be collected to which the model can be fitted. In the final step, the adequacy of the model is tested. The procedure may require much iteration until the fitted equation is determined by the experimenter to be sufficient (Cornell 2002). The most effective choice of model and location of design points is the focus of the experimenter. The best set of points is chosen under the following constraints: (1) the size and shape of the experimental region, (2) the number of desired experimental runs, and (3) the type of model used for

constructing the map of the response (Kettaneh-Wold 1991; Kettaneh-Wold 1992). Most often, the polynomial model is selected to represent the response surface since it can be expanded through a Taylor series to improve accuracy (Cornell 2002). A first or second degree model is usually chosen to represent the surface since it requires fewer observations. Third degree or higher ordered models are seldom utilized. The point estimate forms of the models are listed in Eq. 3.3 and 3.4.

$$y = \beta_o + \sum_{i=1}^u \beta_i x_i \quad (3.3)$$

$$y = \beta_o + \sum_{i=1}^u \beta_i x_i + \sum_{i \leq w} \sum_{w \geq i}^u \beta_{iw} x_i x_w \quad (3.4)$$

In chemical product design, a colinearity effect is usually imposed on the model by Eq. 3.11, which means that the constituent fractions must sum to one and each constituent fraction must lie between zero and one. While this technique does not affect the utilization of the model, it does impact the interpretation of the regression coefficients of the property models. In data-driven approaches, this colinearity affect appears in Scheffe simplex-lattice designs that use canonical models (Eq. 3.14 and 3.15).

$$y = \sum_{i=1}^u \beta_i^* x_i \quad (3.14)$$

$$y = \sum_{i=1}^u \beta_i^* x_i + \sum_{i < w} \sum_{w > i}^u \beta_{iw}^* x_i x_w \quad (3.15)$$

For these types of designs, the location of the response of a mixture made up of exactly zero constituents must be identically zero meaning that the coefficient β_o is zero. Conversely, mixture designs using polynomial models are known as Cox designs where β_o is a non-zero factor used to center the design over the search space. Although the two model types use different parameters,

they predict identical responses (Cornell 2002). The reason for this result is that the regression coefficients represent different entities in each model. In the Scheffe model, the regressed parameters represent combined effects comprised of contributions from pure components, colinear interaction effects, and nonlinear interaction effects. The Cox models remove some of the colinearities in mixture design, but depend on good experimental design controls to limit the effect of other colinearities, resulting in parameters that represent only pure component linear and nonlinear effects. The parameters of the two model types are related according to Eq. 3.17 and 3.18.

$$\beta_i^* = \beta_0 + \beta_i + \beta_{ii} \quad (3.17)$$

$$\beta_{iw}^* = \beta_{iw} - \beta_{ii} - \beta_{ww} \quad (3.18)$$

The estimates of the regressors can be found using least squares regression. To better interpret the impact of regressors on the predicted attribute or physical-chemical property response of the design, it is beneficial to view the design within the property clustering framework.

As discussed in Chapter 6, property clustering is a transformation technique that helps facilitate a reverse problem formulation of a chemical product design. Utilizing property clustering in a reverse problem solving role not only avoids combinatorial explosion, but offers the potential for solving process, mixture, and molecular design problems simultaneously (Eden *et al.* 2003, Solvason *et al.* 2008). In situations where models with the necessary degree of accuracy do not yet exist, chemometric models of attributes ($k \in a$) or physical-chemical properties ($j \in p$) like the Scheffe canonical and Cox polynomial can be utilized and incorporated into the existing property clustering framework while providing two benefits not found in traditional experimental design. First, the experimental design points on which the model is

based are mapped into the property cluster design space and checked against the targets used in the design. This visualization technique is appropriate regardless of the number of components investigated, thereby providing a means for ensuring the design space has been properly explored. The procedure for evaluating the efficiency of coverage is related to how near the optimum is to the design space. It would be unusual to measure this as most designs utilize pure component measurements and involve interpolation. However, should they not, gradient, non-gradient, and specialized optimization techniques can be used to find the optimum and a new mixture design can be conducted at the predicted optimum (Lazic 2004, Brandvik 1998, Brandvik and Daling, 1998).

Second, most experiment based models utilize regression coefficients as estimates of the effects of each property on the response. Depending on the type of regression utilized, the interpretation of these regression coefficients can be quite difficult. Visualizing the problem in property space consolidates each components impact on the mixture, aiding in the ability to screen components. For systems of up to 3 properties, the entire experimental design and all associated regression coefficients can be represented on a single property cluster diagram. For example, a 6 component system can be represented in a single ternary diagram in Fig. 7.1 where the component effects are estimated according to the response model chosen (see Section 7.1.1 and 7.1.2), the feasibility region is estimated as discussed in Chapter 6, and the design points are estimated according to Section 7.1.3. Additional properties are either represented with additional diagrams or solved algebraically. Consolidating the numerous effects of the various components on the response into distinguishable values may be used to surmise which of the components has the largest effect and thus is most important.

Third, constraints on the applicable range of the model and the potential mixtures that are formed can be defined using Hotelling's T^2 distribution or the Bonferroni approximation on the predictive power of the models (Johnson and Wichern 2007). These can be used to apply a second sink region which must also be matched simultaneously with the target feasibility sink region (Solvason *et al.* 2008; Solvason *et al.* 2009).

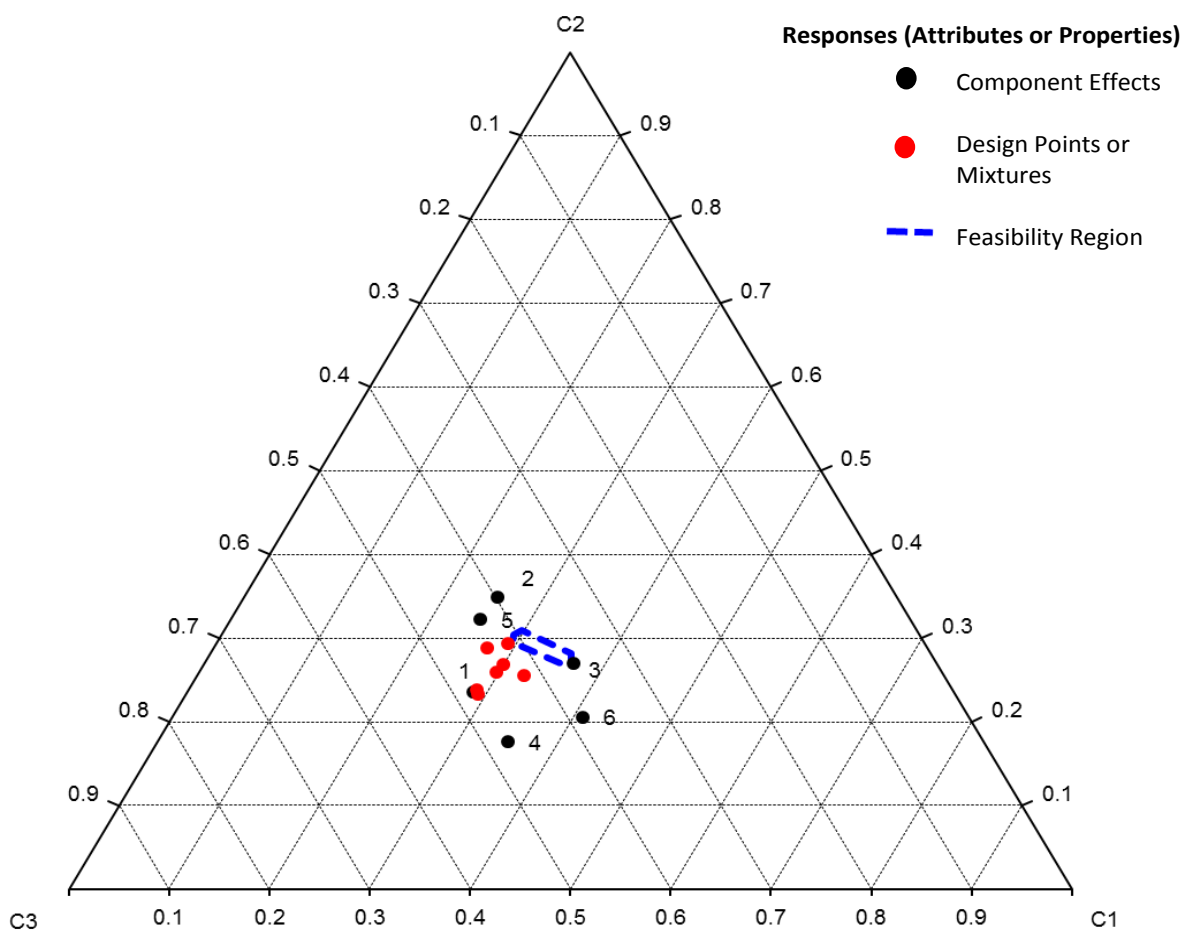


Figure 7.1: The Component Effects, Experimental Design Points, and Feasibility Region of a Six-component Mixture Design Mapped to the Property Cluster Domain.

One difference between the use of regression based property models and conventional property models is that the flexibility of parameter estimation allows for the use of negative regressors, especially when the parameters are highly confounded. If the coefficients only

represented pure component values, then the property would always be positive, but if a regressor is negative, then it follows that coefficient is indicative of a strong colinearity and/or nonlinearity effect overwhelming the weak, positive linear effect. This situation can cause problems within the traditional property clustering framework, but is not unsolvable.

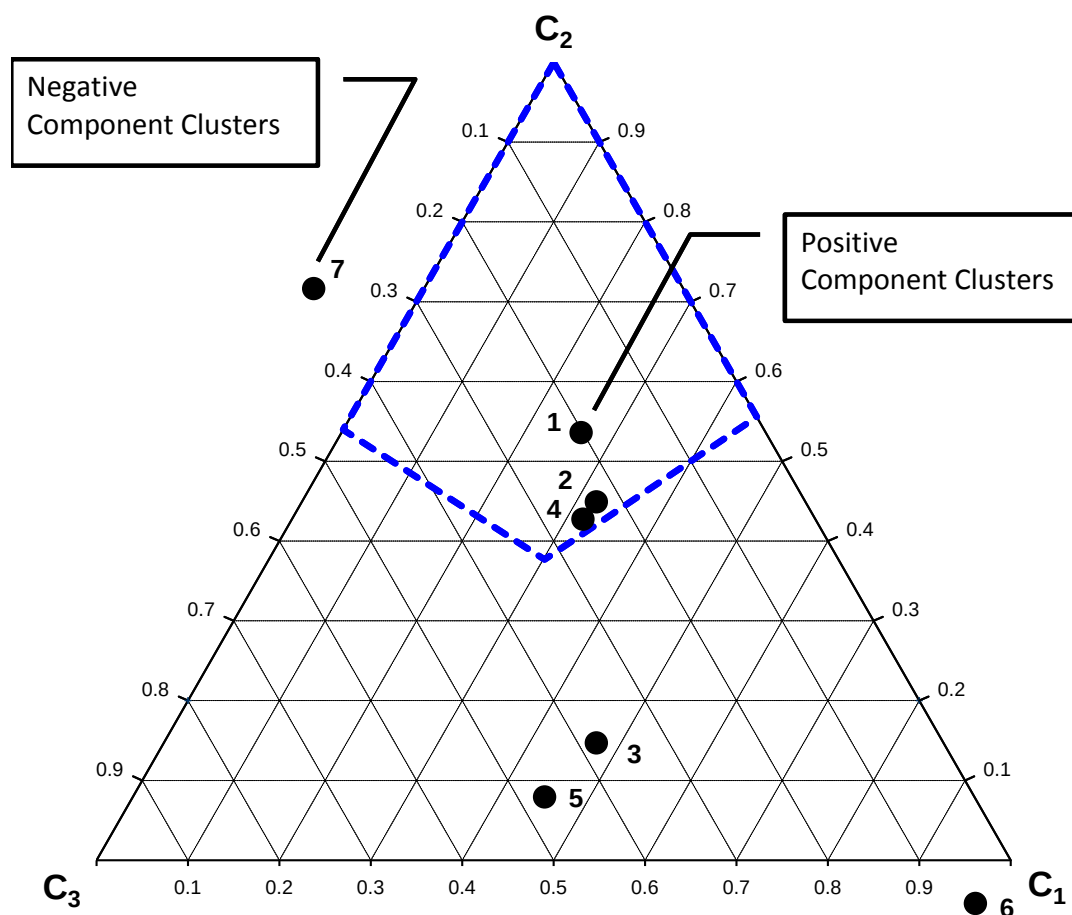


Figure 7.2: A Property Cluster Simplex Diagram Showing a Seven-Component System using Property Operators Derived from an Acetaminophen Excipient Design (Martinello *et al.* 2006; Solvason *et al.* 2008). Clusters 1, 2, and 3 Represent the Properties Repose Angle, Water Content, and Compressibility.

For example, in Fig. 7.2 a mixture design of seven components is shown in a ternary cluster diagram. Components 6 and 7 both had component effects consisting of negative regression coefficients which, when clustered and mapped using Cartesian coordinate transformations of Eq. 6.25 and Eq. 6.26, place them outside the traditional cluster diagram.

Hence, it is possible that when using models comprised of regression coefficients in CAMD, some solutions may require mixing with chemical constituents outside of the cluster space defined by the ternary diagram. As a result, it is necessary to define the region outside of the cluster space as negative cluster space and the region inside the ternary simplex as positive cluster space.

Negative cluster space can consist of a multitude of regions. The total number of negative cluster regions, N_R , is a function of the number of p properties or a attributes, depending on the system in question (Solvason *et al.* 2008; Solvason *et al.* 2009). For a property system, the following holds:

$$N_R = 2p \text{ or } N_R = 2a \quad (7.1)$$

For a $p = 3$ or $a = 3$ property or attribute solution the negative cluster space is comprised of six distinct regions as shown in in Fig. 7.3. These regions are of two types: those with clusters greater than one are called Type I regions and those with negative clusters are called Type II regions (Solvason *et al.* 2008; Solvason *et al.* 2009).

$$C_{ji} > 1 \text{ or } C_{ki} > 1, \text{ Type I Regions} \quad (7.2)$$

$$C_{ji} < 0 \text{ or } C_{ki} < 0, \text{ Type II Regions} \quad (7.3)$$

These regions are related to one another since in the derivation of property clusters, Eden *et al.* (2003) notes that the clusters must sum to one, which also holds for clusters of attributes.

$$\sum_{j=1}^p C_{ji} = 1 \text{ or } \sum_{k=1}^a C_{ki} = 1 \quad (7.4)$$

For the three property example Eq. 7.4 implies that two negative property clusters equate to a third cluster with a value greater than one. Likewise, two clusters with a value greater than one equate to a third cluster with a value less than one. The number of different types of (positive and negative) cluster regions N_T is related to the number of properties evaluated according to Eq. 7.16 (Solvason *et al.* 2008; Solvason *et al.* 2009).

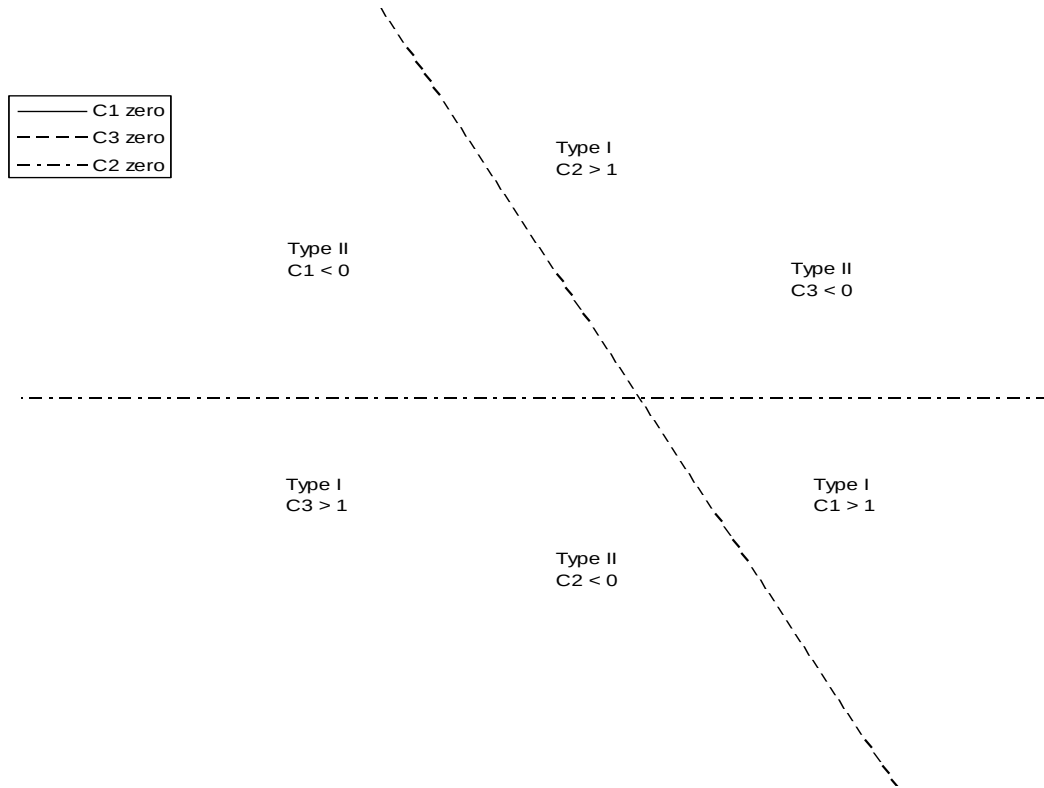


Figure 7.3: Details of Negative Property Cluster Space for a Three-Property System (Solvason *et al.* 2008; Solvason *et al.* 2009).

$$N_T = p \text{ or } N_T = a \quad (7.5)$$

Property clusters in the positive cluster regions may also be estimated from three negative clusters and three clusters greater than one, respectively. These values may only be used to

ascertain effects and may not be used in mixing without violating the monotonically increasing rule for mixing proposed by Shelley and El-Halwagi (2000) and Eden *et al.* (2003).

To discuss this limitation it is noteworthy when dealing with negative property clusters to first investigate the constraints on the Augmented Property Index (*AUP*). The non-dimensionalized properties are summed to create the *AUP*. While a negative regression coefficient may create a negative property operator, the advent of such constructs must be constrained by the following rule (Solvason *et al.* 2008; Solvason *et al.* 2009):

Rule 5. All *AUP* values of the components must be positive.

The rule can be written as equation 7.6

$$AUP > 0 \tag{7.6}$$

In order to meet Rule 4 it is often necessary to adjust the reference property values in Eq. 6.4. Although the adjusted references give different values for the clusters, the underlying property values are unchanged. For example, the cluster diagram shown in Fig. 7.4 is altered by changing the reference values for the property operators. Although the clusters of the three components reside in a different location, their relative proximity to each other and the feasibility region remains the same. In other words, for the example shown in Fig. 7.4, component 3 will always be closest to the feasibility region while component 1 will always be farthest from the feasibility region regardless of the property operator reference chosen. Furthermore, when the experimental design points A, B, C, D, and E from the adjusted cluster diagram are transformed back to component space, the resulting solutions are identical to the design points derived using the original references. This is an important aspect of property clustering as the flexibility to adjust the property operator reference values allows negative regressors to be utilized. An algorithm that satisfies Rule 4 and Eq. 7.6 was presented in Chapter 6.

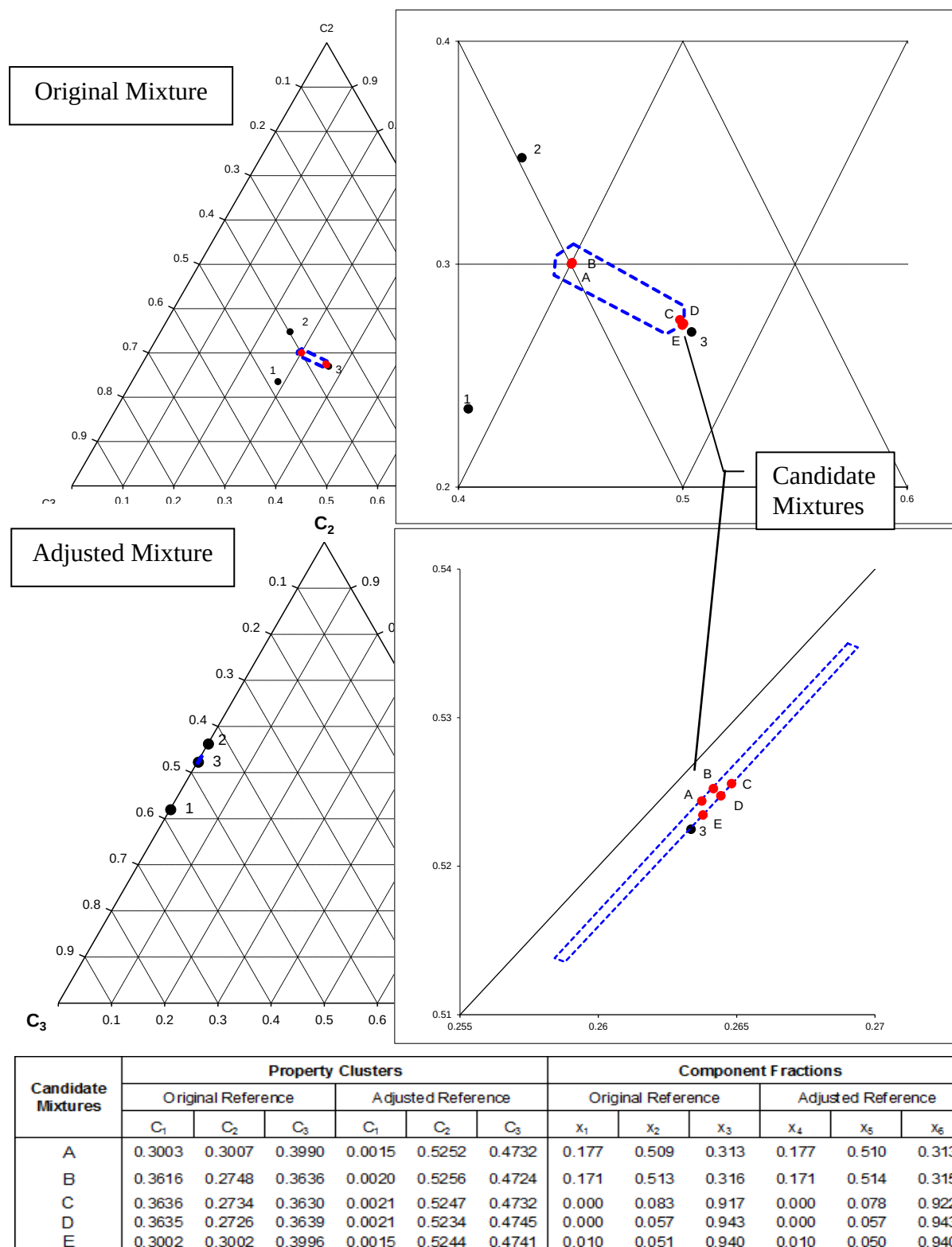


Figure 7.4: A Three Component System of Polyethylene, Polystyrene, and Polypropylene Mapped to Cluster Space using Different Property Operator Reference Values.

7.1.1.1. Canonical Models

The simplest and most effective regression model to use with design of experiments is the Scheffe canonical model of Eq. 3.3. Using the regression coefficients as estimates of the effects of a component on the overall mixture, Eq. 3.3 can be transformed into the property operator model of Eq. 6.1 for $k \in a$ attributes using the nondimensionalized property operator expression shown in Eq. 7.7:

$$\Omega_{ik} = \frac{\beta_{ik}^*}{\psi_k^{ref}(y_k)} \quad (7.7)$$

The procedure for clustering the property operator expression then follows the method introduced in Section 6.1. It should be noted that the Scheffe canonical model ensures an efficient CAM^bD computation to the detriment of the interpretation of its parameters, which may limit its use in secondary optimization schemes that use parameter interpretation to predict new experiments (e.g. genetic algorithm). The reason for this result is that the regression coefficients for the Scheffe models are combinations of linear, collinear, and nonlinear effects as shown in Eq. 3.17 and often result in negative regressors and clusters. However, as long as the *AUP* of pure component effects remain positive, mixtures generated from the negative clusters will still meet the inter-conservation and intra-conservation rules of property clustering.

7.1.2. Polynomial Models

The approach in Section 7.1.2 is valid for setting up regression based models for mixing and developing lists of candidate solutions. However, if the objective is to screen constituents, then some knowledge of the pure component effects will be needed in order to reduce additional experimentation each time a new constituent is added to the list of candidates. Since the objective of the screening design is to understand the effects of the constituents on the mixture, then the component clusters need to represent the pure property values void of colinearities and nonlinearities. One method for removing the colinearities is to utilize the Cox modifications on the Scheffe canonical models to create a reference mixture, or standard mixture. Shown in Fig. 3.2 is a simplex diagram published by Cornell (2002) with a standard mixture s and a mixture x with a larger proportion of constituent x_i . Noting that x lies on a line from s to the x_i vertex, then the ratio of the other $u-1$ constituents are in the same relative proportions as the standard mixture. As Smith and Beverly (1997) point out, the Δ_i gradient, or change in response per unit change in x_i , at s along the Cox-effect direction is called the *effect of x_i* , provided x_i is free to range from 0 to 1. The relationship between the mass fraction, standard mixture and gradient is written as Eq. 7.8, which, when inserted into the polynomial expression of Eq. 3.3, results in Eq. 7.9:

$$x_i = s_i + \Delta_i \quad (3.19)$$

$$y(x) = y(s) + \sum_i^u \left(\frac{\beta_i}{1 - s_i} \right) \Delta_i \quad (3.21)$$

where $y(x)$ is the expected response at design point x and $y(s)$ is the expected response at the standard reference mixture. For an attribute system ($k \in a$), it has been shown that the response of the standard mixture $y(s)$ is equivalent to the centering function provided by the regressor β_o (Cornell 2002):

$$\Omega_k^s = \frac{\beta_{ko}}{\psi_k^{ref}(y_k)} \quad (7.8)$$

where Ω_k^s is the property operator for the standard mixture. The second term in Eq. 3.21 provides a direct link between the regressors and the position of the design points to the reference mixture (Cornell, 2002) such that

$$\sum_i^n \left(\frac{\beta_i}{1-s_i} \right) \Delta_i = \sum_i^n \beta_i x_i = y^z \quad (7.9)$$

$$y(x) = y(s) + y^z \quad (7.10)$$

where $y(s)$ is the response at the standard mixture and y^z is the pseudo property value that represents the attributes, k , contribution to the mixture. The equivalent property operator expression for the k^{th} attribute effect or “pseudo” component then becomes Eq. 7.11:

$$\Omega_k^z = \frac{\beta_k}{\psi_k(y_k)_{ref}} \quad (7.11)$$

with the normalized property operator expression of Eq. 7.10, thus results in Eq. 7.12.

$$\Omega_k = \Omega_k^s + \Omega_k^z \quad (7.12)$$

Likewise the augmented property index is rewritten as Eq. 7.13:

$$AUP = AUP^s + AUP^z \quad (7.13)$$

where the AUP^s is the AUP of the standard mixture:

$$AUP^s = \sum_{k=1}^a \Omega_k^s \quad (7.14)$$

and AUP^z is the AUP of the pseudo component effect.

$$AUP^z = \sum_{k=1}^a \Omega_k^z \quad (7.15)$$

Redefining the property clusters in terms of the pseudo property cluster C_k^z and the standard property cluster C_k^s gives equations 7.16 and 7.17.

$$C_k^z = \frac{\Omega_k^z}{AUP^z} \quad (7.16)$$

$$C_k^s = \frac{\Omega_k^s}{AUP^s} \quad (7.17)$$

The sum of these clusters will not give the true cluster, C_k , as defined in Eq. 7.4, without correcting for the different AUP values. This is done using a set of correction factors as shown in Eq. 7.18 and 7.19.

$$F^z = \frac{AUP^z}{AUP} \quad (7.18)$$

$$F^s = \frac{AUP^s}{AUP} \quad (7.19)$$

F^z is the pseudo correction factor and F^s is the standard correction factor. These are combined in Eq. 7.20 to give the relationship between the true property cluster, the pseudo property cluster, and the standard property cluster.

$$C_k = F^s C_k^s + F^z C_k^z \quad (7.20)$$

It should be noted that while the transformation of the original polynomials removes the primary colinearity introduced by Eq. 3.11, it leaves the secondary colinearities such as those introduced by constraints on the constituent ranges. So, although both Scheffe canonical and Cox polynomial models provide estimates of the pure component effects of the chemical

constituents, they are, in fact, not the property or attribute values of the pure components. Kettaneh-Wold (1992) suggests that the best solution maybe to refrain from interpreting the coefficients and rely on the predictions only but notes this solution is not acceptable in practice since the interpretation of regression coefficients is a necessity when the objective is to find component effects in screening situations. One method to handle the estimate of the effects is to constrain the solutions to the CAM^bD within a feasibility region where the estimates of the effects are considered valid, known as a model feasibility region (MFR). For multivariate normal responses $N_p(\mu, \sigma)$, a component can be tested against the MFR using Hoetelling's T^2 .

$$T^2 = n(\bar{Y} - Y_e)' S^{-1} (\bar{Y} - Y_e) \quad (7.21)$$

where n is the number of experiments (i.e. mixtures), $\bar{Y}$ is a vector of estimated means of the response (e.g. $Y = A$ when the responses of interest are attributes and $Y = P$ when the responses are physical-chemical properties), and Y_e is a vector of responses of a test mixture e . S is a $p \times p$ (or $a \times a$ for attribute responses) matrix of the estimated variance and covariance between the responses.

$$S = \frac{1}{n-1} \sum_{e=1}^n (Y_e - \bar{Y})(Y_e - \bar{Y})' \quad (7.22)$$

For the $N_p(\mu, \sigma)$ population, the null hypothesis H_0 is that the test mixture is described by the response model and that the alternative hypothesis H_1 is that it is not.

$$H_0: \mu = Y_e \quad (7.23)$$

$$H_1: \mu \neq Y_e \quad (7.24)$$

Hoetelling's T^2 is distributed as the following:

$$T^2 \sim \frac{(n-1)p}{n-p} F_{p, n-p}(\xi) \quad (7.25)$$

where $F_{n, n-p}(\xi)$ is the single tailed F distribution. The null hypothesis is rejected if T^2 calculated from Eq. 7.21 is greater than the critical T^2 calculated from Eq. 7.25, as shown in Eq. 7.26.

$$\text{Reject } H_0 \text{ if } T^2 > \frac{(n-1)p}{n-p} F_{p, n-p}(\xi) \quad (7.26)$$

For the existing chemical constituents present in the training set of an aCAM^bD, confidence intervals on the mean responses can be derived from Eq. 7.21 (Johnson and Wichern). However, since the primary objective of the CAM^bD is to generate new mixtures, it is prudent to reconstruct Eq. 7.21 in terms of the difference between the predicted response and the actual response:

$$T^2 = \left(\frac{B'x_e - Y_e}{\sqrt{x_e'(X'X)^{-1}x_e}} \right)' \left(\frac{n}{n-u-1} S^E \right)^{-1} \left(\frac{B'x_e - Y_e}{\sqrt{x_e'(X'X)^{-1}x_e}} \right) \quad (7.27)$$

where x_e is the test set of new mixtures, X is the set of existing mixtures, $B'x_e$ are the predicted responses of the set of test mixtures, Y_e is the measured response of the test mixtures, n is the total number of experiments in the training set, u is the total number of chemical species in the mixture, and S^E is the unexplained variance in the fitted model. Using multiple linear regression, the confidence intervals of each test mixture for each k attribute (or j property) can be written as Eq. 7.28 (Johnson and Wichern 2007):

$$Y_{ke} \pm \sqrt{\left(\frac{a(n-u-1)}{(n-u-a)} F_{a, n-u-a}(\xi) \right) \cdot \left(\frac{n}{n-u-1} \right) \cdot S_{kk}^E \cdot x_e'(X'X)^{-1}x_e} \quad (7.28)$$

The $k \in a$ attribute boundaries of the MFR are calculated by applying Eq. 3.3 (or Eq. 3.14) and Eq. 7.28 to each experiment in the training set and taking the maximum and minimum of the responses. The cluster feasibility region is then built using the same procedure outlined in Chapter 6. Further analysis of this region is given in Chapter 8.

In some situations the property, attribute, and/or chemical constituents are highly confounded with large amounts of covariance. For this situation, Kettaneh-Wold (1992) suggests the use of decomposition techniques like Principal Component Regression (PCR) and Partial Least Squares on to Latent Surfaces (PLS) to find underlying, orthogonal properties or components that can significantly reduce the complexity of the problem. These techniques were discussed in detail in Chapter 3 and are applied to aCAMD with clusters in Chapter 9.

7.2. Attribute Clustering Algorithms

The attribute clustering algorithm for an attribute based computer aided mixture design (aCAM^bD) proceeds in the same manner as CAM^bD with the some additions. First, regardless of the model type chosen, a training set of validated experimental design points (i.e. mixtures) are available to for testing against the feasibility region. These mixtures can be calculated directly from Eq. 6.6 and, if $a = 3$ can be plotted by the procedure outlined in Chapter 6. Second, if a Cox polynomial model is chosen for design, a pseudo feasibility region must be calculated from the feasibility region limits adjusted for the standard reference mixture by rearranging Eq. 7.12, 7.13, and 7.20. The same holds true for the MFR limits. An outline of the clustering algorithm is shown in Fig. 7.5.

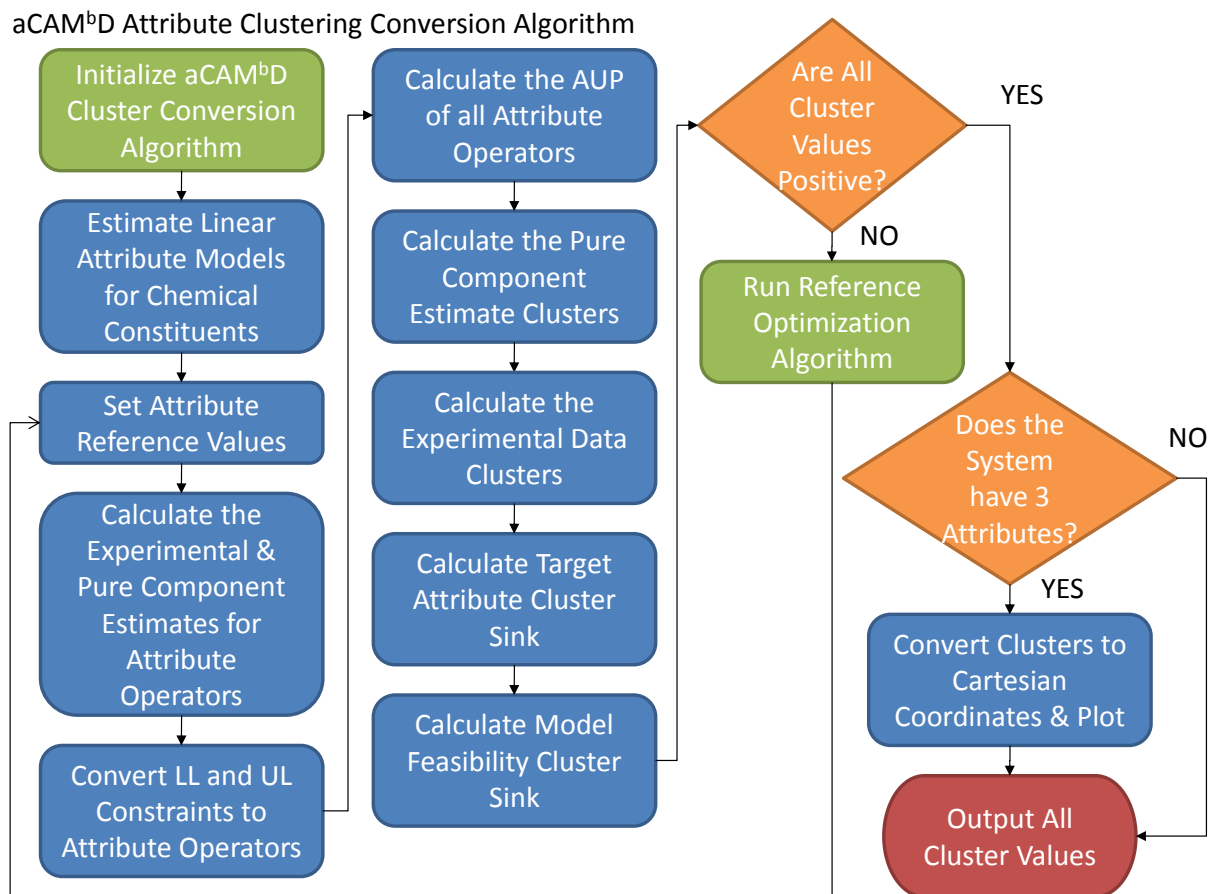


Figure 7.5: Clustering algorithm for Attribute based Computer Aided Molecular Design (aCAM^bD)

The choice of regression model (canonical or polynomial) dictates the specific equations utilized in the clustering approach as shown in Table 7.1 and 7.2.

Table 7.1. Attribute Clustering Procedure for Scheffe Canonical Models

Step	Description	Equation
1	Select Linearized Canonical Model	3.14, 3.15
2	Fit Model to Experiment Design Points and Responses to Determine Models Parameters	3.10
3	Guess Property Operator Reference Values and Calculate Pure Component Effect Property Operator Values	7.7

4	Convert UL and LL Feasibility Constraints to Property Operators	6.7
5	Calculate the AUP of the Pure Component Effect Property Operators and the Feasibility Constraint Property Operators	7.7, 6.7, 6.5
6	Calculate the Pure Component Effect Clusters	6.6
7	Calculate the Feasibility Region Sink	6.6, 6.7
8	Calculate the Variance, Covariance, and Confidence Intervals of the Experimental Design	7.20
9	Calculate the Model Feasibility Region using Confidence Intervals around the Responses of the Experimental Design Points	7.28, 6.6, 6.7
10	Run Reference Optimization Algorithm for Non-Positive Clusters	6.8-6.12
11	Convert all Clusters to Cartesian Coordinates and Plot	6.23-6.26

Table 7.2. Attribute Clustering Procedure for Cox Polynomial Models

Step	Description	Equation
1	Select Linearized Polynomial Model	3.3, 3.4
2	Fit Model to Experiment Design Points and Responses to Determine Models Parameters	3.10
3	Guess Property Operator Reference Values and Calculate Pseudo Pure Component Effect and Standard Property Operator Values	7.8, 7.11
4	Convert UL and LL Feasibility Constraints to Pseudo Property Feasibility Constraints	6.7, 7.12
5	Calculate the UL and LL of the Pseudo Feasibility Region AUP	7.15
7	Calculate the AUP of the Pseudo Pure Component Effect and Standard Reference Mixture	7.14, 7.15
8	Calculate the Pseudo Component Effect and Standard Reference Clusters	7.16, 7.17
8	Calculate the Variance, Covariance, and Confidence Intervals of the Experimental Design	7.20
9	Calculate the Model Feasibility Region using Confidence Intervals around the Responses of the Experimental Design Points	7.28, 6.6, 6.7
10	Run Reference Optimization Algorithm for Non-Positive Clusters	6.8-6.12
11	Convert all Clusters to Cartesian Coordinates and Plot	6.23-6.26

7.3. aCAM^bD using Experimental Data and Data Driven Models

Once the attribute (or property) clusters have been calculated, aCAM^bD can proceed in a similar manner to CAM^bD. Since Shelley and El-Halwagi (2000) have shown that a property cluster is conserved through mixing, the fractional calculation of any mixture in the cluster design space can be found. For Scheffe canonical models, the property operator expression is identical to Eq. 6.1 and the design procedure follows the outline provided in Section 6.2 with some additional rules. First, immediately after testing for Rule 1, the mixtures should be tested against the MFR using Rule 6, discarding any solutions that fail the test as infeasible.

Rule 6. Cluster values of the source (or mixture of sources) must be contained within the MFR of the sink in the cluster domain.

Likewise, following the tests on Rule 2/3, the mixture should be tested against Rules 7/8 for suitable MFR *AUP* values.

Rule 7. The values of the Augmented Property Index (*AUP*) for the source or mixture of sources must fall within the *AUP* range of the MFR.

Rule 8. The *AUP* of the candidate must match the *AUP* of the MFR sink at the candidate point as calculated from the sink boundary points.

For Cox polynomial models, the calculation of the mixture fractions requires the rewriting of Eq. 6.18 in terms of a mixture of pseudo properties and standard properties. First Eq. 6.1 is written as Eq. 7.29.

$$\Omega_{kM} = \sum_i^u \Omega_{ki} \cdot x_i \quad (7.29)$$

The normalized property operator for each component effect can be written as Eq. 7.12 and inserted into Eq. 7.29 to give Eq. 7.30

$$\Omega_{kM} = \sum_i^u \Omega_{ki}^z \cdot x_i + \Omega_k^s \quad (7.30)$$

Dividing the expression by the *AUP* of the mixture and inserting standard mixture correction factor and cluster definition gives Eq. 7.31.

$$C_{kM} = \frac{\sum_i^u x_i \Omega_{ki}^z}{AUP_M} + \frac{\Omega_k^s}{AUP_M^s} \cdot F_M^s \quad (7.31)$$

Inserting the cluster definition of Eq. 7.17 into Eq. 7.31 and rearranging gives Eq 7.32.

$$C_{kM} - F_M^s C_{kM}^s = \frac{\sum_i^u x_i \Omega_{ki}^z}{AUP_M} \quad (7.32)$$

Noting that the left hand side of Eq. 7.32 is the same as the product of the pseudo correction factor and the pseudo property cluster of the mixture, the equation is rewritten as Eq. 7.33.

$$F_M^z C_{jM}^z = \frac{\sum_i^n x_i \Omega_{ji}^z}{AUP_M} \quad (7.33)$$

Inserting the pseudo cluster definition to remove property operator in favor of the cluster and rewriting the *AUP* of the mixture in pseudo terms gives Eq. 7.34:

$$C_{jM}^z = \frac{\sum_i^n x_i AUP_i^z C_{ji}^z}{AUP_M^z} \quad (7.34)$$

Eq. 7.35 assumes a familiar form of the relative cluster arm using the pseudo relative cluster arm δ^z as defined by Eq. 7.36.

$$C_{jM}^z = \sum_i^n \delta_i^z C_{ji}^z \quad (7.35)$$

$$\delta_i^z = \frac{x_i AUP_i^z}{AUP_M^z} \quad (7.36)$$

The pseudo relative cluster arm maintains the monotonically increasing criteria imposed by Eden *et al.* (2003) provided that all of the augmented property indices used in the solution of the problem are positive; a constraint placed on the solution by Eq. 7.6. A negative *AUP* would violate this relationship and prevent the proper cluster solution from being obtained. The relative cluster arms of the pseudo mixing rule in Eq. 7.35 are indicative of a mix involving a pseudo feasibility region. The pseudo feasibility region is defined in the same manner as the true feasibility region but corrects for the standard mixture property values using Eq. 7.12. Often this means that the pseudo region is located in the negative cluster space. The resulting pseudo relative cluster arms represent the addition of the pseudo components to move the mixture into this region, only now each pseudo component better represents its contribution to the mixtures' properties, void of most colinearities. For example, suppose a $u = 3, p = 3$ component system is being evaluated by property clustering and that parameters for both the Cox polynomial and Scheffe canonical models are estimated. As shown in Fig. 7.6, identical mixtures A, B, C, D, and E are calculated using both the conventional cluster mixing approach with Scheffe models or the pseudo clustering approach using Cox models. This is expected since the normal Scheffe and Cox mixing designs achieve the same property response plots (Cornell 2002). It is the author's recommendation that for ease of use, the traditional property clustering method should be used

with the Scheffe property operators, while the pseudo property clustering method using Cox property operators should be reserved for interpretation of component effects on the mixture properties. An interesting benefit of using pseudo property clustering with Cox models over the traditional method is the ability to visualize the components impact on the mixtures properties simultaneously. In the traditional techniques, each response plot is mapped onto the component space. Ignoring the combinatorial explosion issue for a moment, it can be seen that when two or more iso-properties are parallel or conflict in direction, it can be difficult to know where to move the mixture visually. This problem is compounded exponentially when multiple components are evaluated; leading researchers to either limit the number of components in the experiment or use powerful statistical techniques such as PLS. Pseudo property clustering offers a medium ground between the two methods and in some cases can be used in conjunction with PCR and PLS to further clarify solutions, especially when performing screening designs. Rules governing the interpretation of the cluster points in the property cluster space are listed as follows:

Rule 9. The visual distance from the standard mixture to a component cluster point is indicative of the relative magnitude of the components effect on the response.

Rule 10. If the constituents lie on opposite sides of a line which passes through the standard reference mixture, then the constituents are said to be inversely related.

A flowchart of the aCAM^bD method is shown in Fig. 7.7 and Table 7.3 lists the equations used in the design.

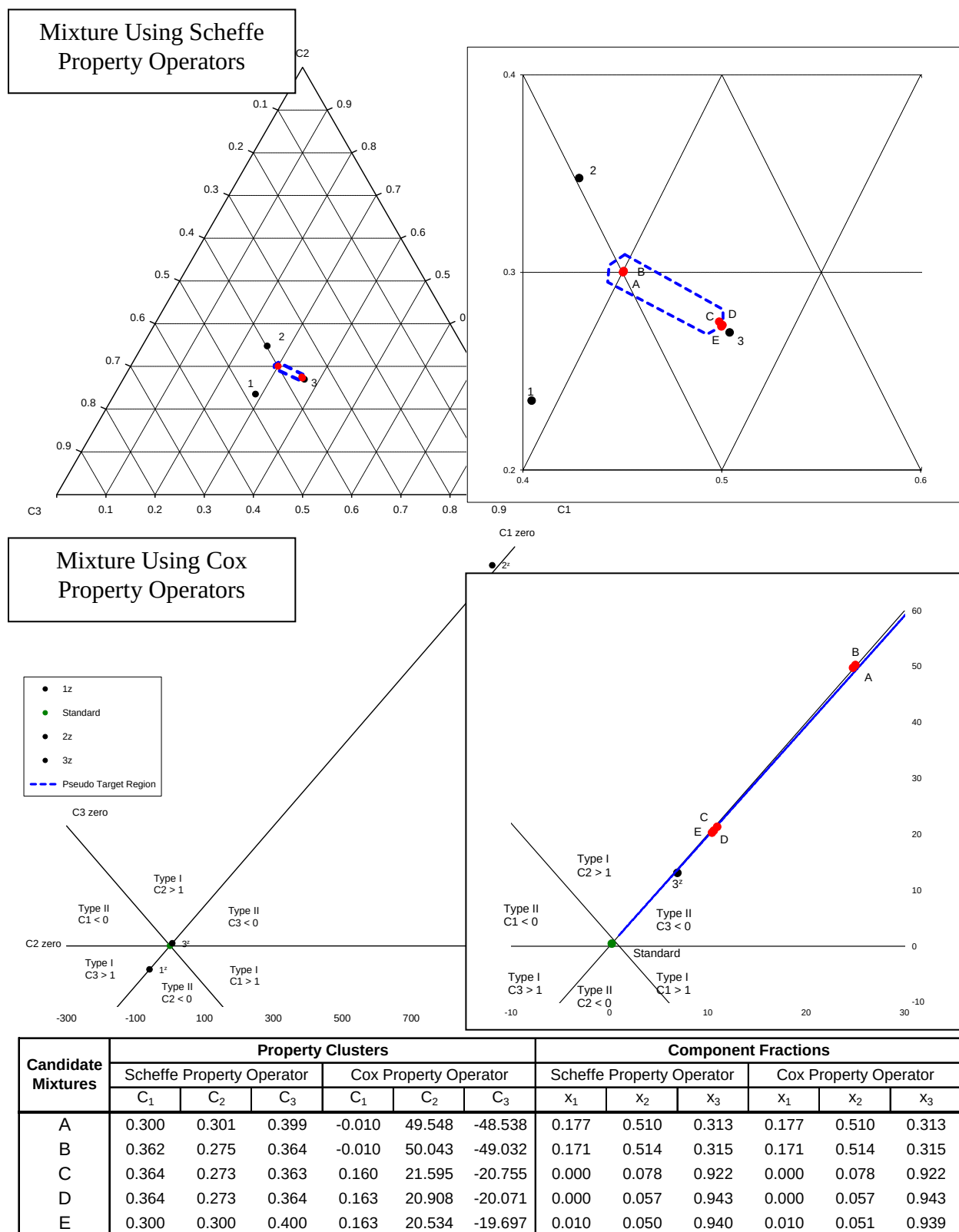


Figure 7.6: A Three Component System Mapped to Cluster Space using both Scheffe and Cox Property Operator Models.

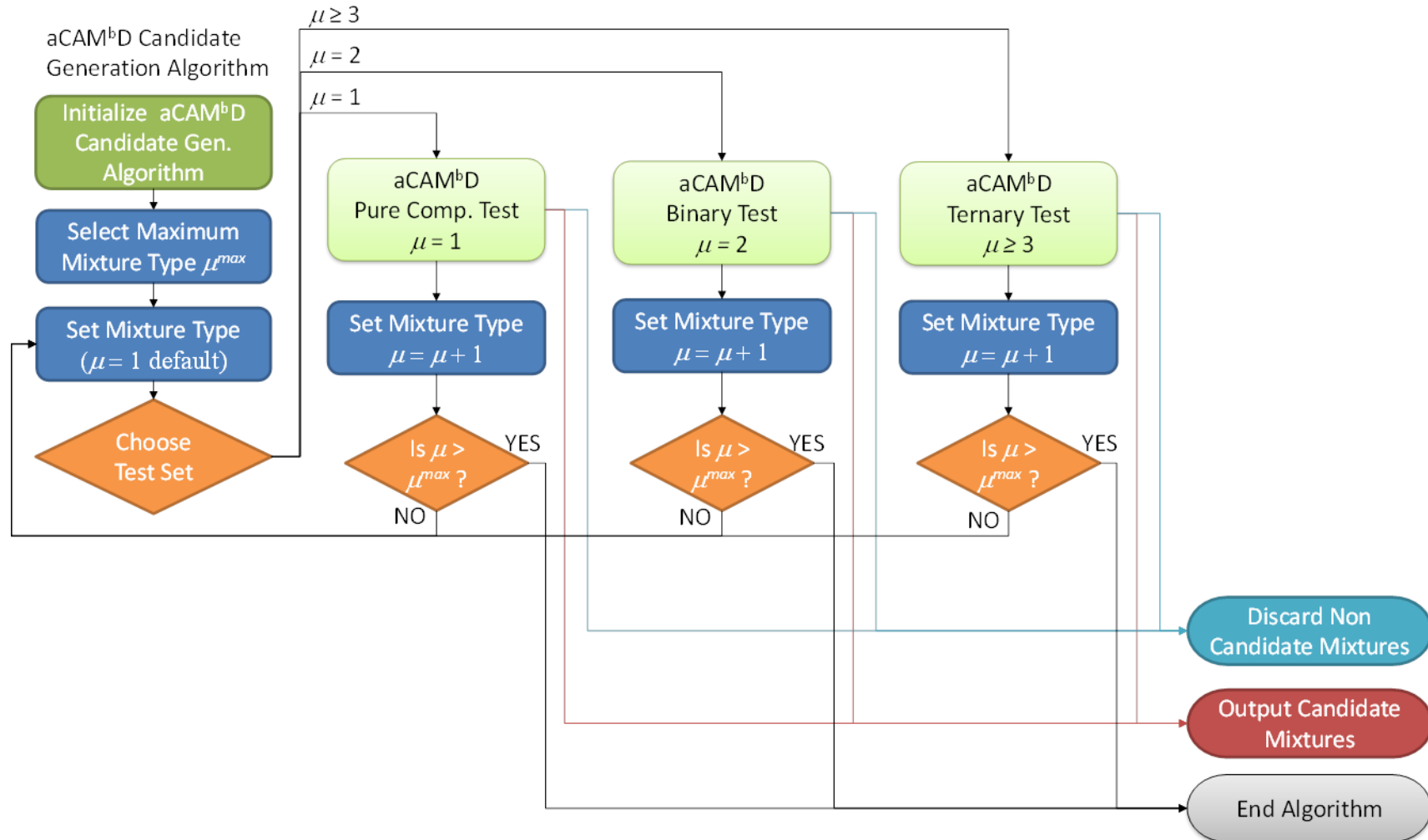


Figure 7.7: Overview of the Attribute Based Computer Aided Mixture Design (aCAM^bD) Method.

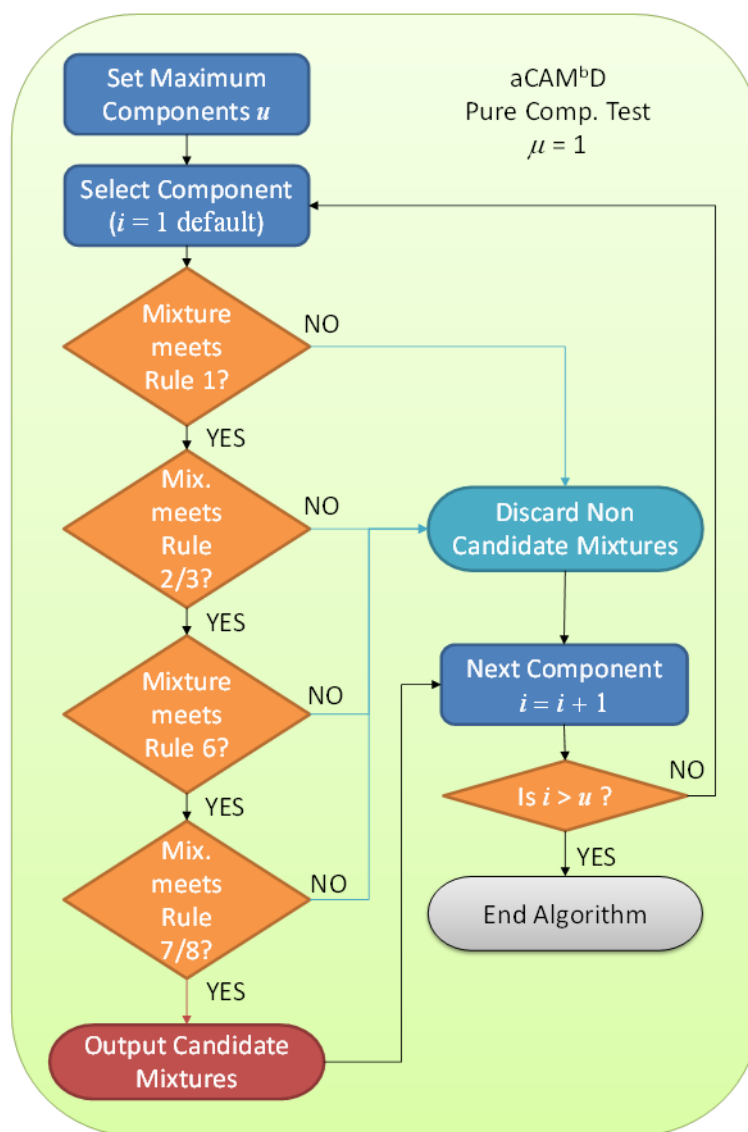


Figure 7.8: aCAM^bD with Clusters test on Pure Components

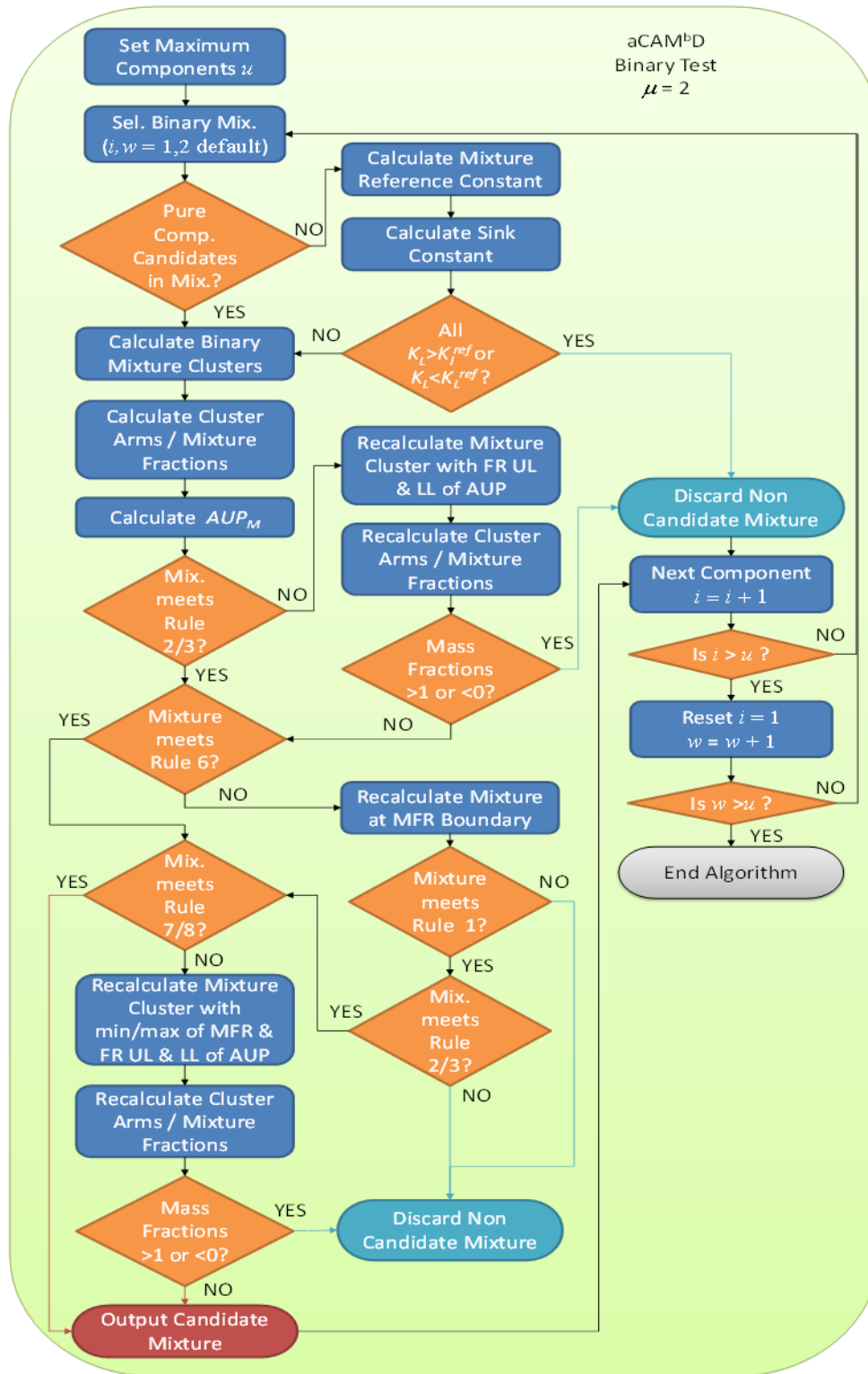


Figure 7.9: aCAM^bD with Clusters test on Binary Mixtures

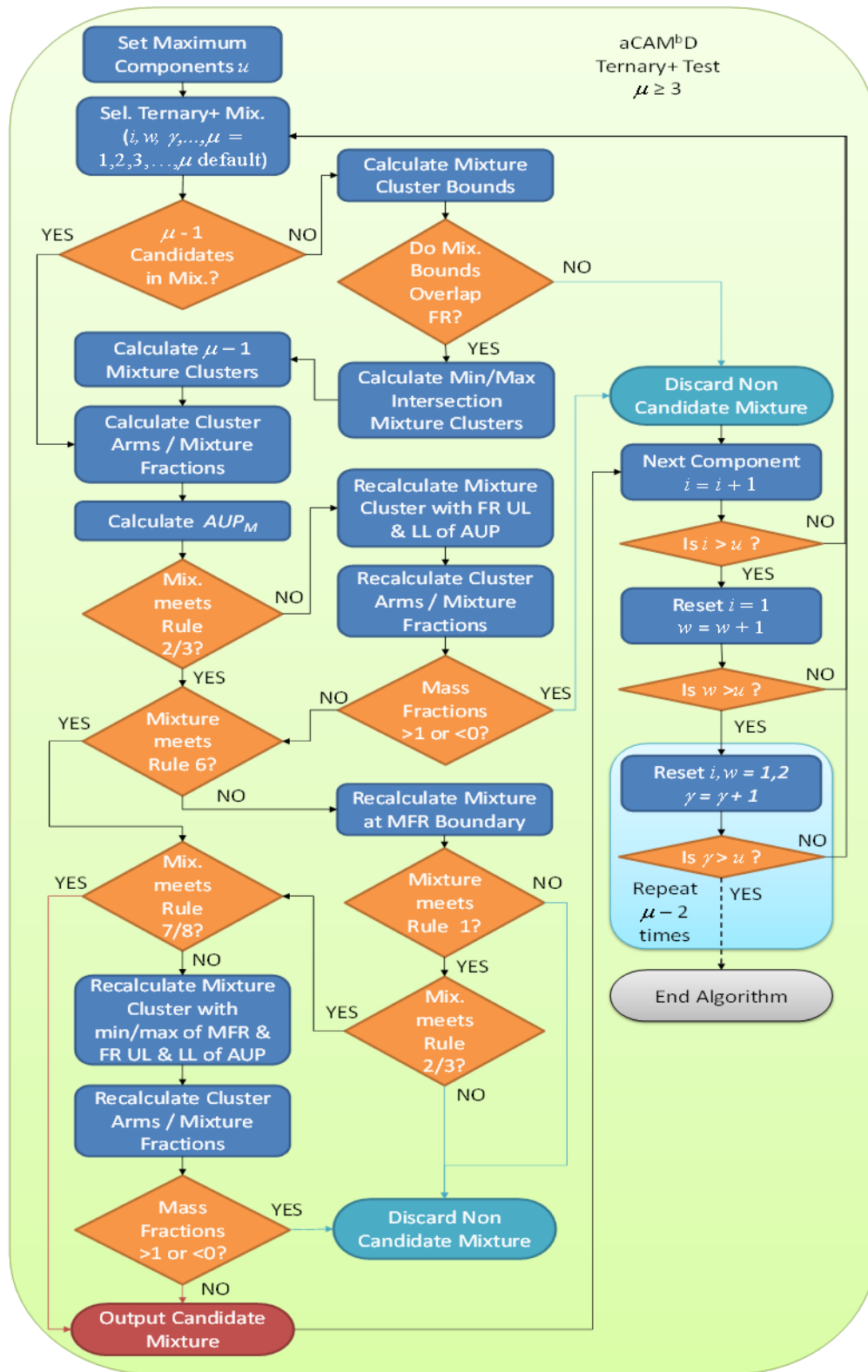


Figure 7.10: aCAM^bD with Clusters test on Ternary and Larger (Ternary+) Mixtures

Table 7.3: The Attribute Based Computer Aided Mixture Design (aCAM^bD) Method.

Step	Description	Equation
1	Set μ^{max} Maximum Dissimilar Species in Mixture	-
2	Discard Pure Component Clusters (from aCAM ^b D Conversion Algorithm) that Fail Rule 1	6.6, 6.7
3	Discard Pure Component Clusters that Fail Rule 2/3	6.5, 6.7
4	Discard Pure Component Clusters that Fail Rule 6 and reside outside the Model Feasibility Region	7.25, 7.26, 6.6, 6.7
5	Discard Pure Component Clusters that Fail Rule 7/8, Output Pure Component Candidates	6.5, 6.7, 7.25, 7.26
6	For Binary Mixtures with Pure Component Candidates, Calculate the Clusters at the Intersections of the Sink and the Mixtures	6.31, 6.23, 6.24
7	Calculate the Cluster Arms	6.21, 6.22
8	Calculate the Fractions of Each Pure Component	6.32
9	Calculate the AUP of Mixtures	6.20
10	For Mixtures that Fail Rule 2 or 3, set the AUP _M at the UL or LL of the Sink and Recalculate the Component Fractions, Discard Binary Mixtures with Component Fractions < 0 or > 1	6.20
11	For Mixtures that Fail Rule 6, Recalculate the Mixture at the MFR Boundary, Discard Solutions that Fail Rule 1 and/or Rule 2/3	6.31, 6.21, 6.32
12	For Mixtures that Fail Rule 7/8, set the AUP _M at the Minimum UL or Maximum LL of the both the FR and MFR Sinks and Recalculate the Component Fractions, Discard Binary Mixtures with Component Fractions < 0 or > 1, Output Candidate Mixtures	7.25, 7.26, 6.6, 6.7, 6.31, 6.21, 6.32
13	For Binary Mixtures with no Pure Component Candidates, Calculate the Reference Constant for each Mixture	6.23 6.24, 6.30
14	Calculate the Sink Constants for each Mixture, Discard Mixtures where all Sink Constants are > or < Reference Constant	6.23-6.24, 6.27-6.29
15	Calculate the Clusters at the Intersections of the Sink and the Mixture and Repeat Steps 7-14, Output Candidate Mixtures	6.31, 6.23, 6.24
16	For $\mu \geq 3$ Ternary+ Mixtures with $\mu - 1$ Candidates, Calculate the $\mu - 1$ Equivalent Clusters	6.31, 6.23, 6.24
17	Calculate the Clusters at the Intersections of the Sink and the $\mu \geq 3$ Mixture and Repeat Steps 7-12, Output Candidate Mixtures	6.31, 6.23, 6.24
19	For $\mu \geq 3$ Ternary+ Mixtures with no $\mu - 1$ Candidates, Calculate the Cluster Range of the Pure Components	6.34, 6.35, 6.6

20	Calculate the Cluster Range of the Feasibility Sink, Discard Mixtures whose Pure Component Cluster Range is > Sink Cluster Range	6.34, 6.35, 6.6, 6.7
21	Repeat Steps 16-20 until μ^{max} is reached	-

After the initial aCAM^bD has been conducted, it is useful to design a feedback loop to help drive the design toward an optimal mixture. The determination of optimal conditions from an experimental design is among the most complex problems for a researcher (Lazic 2004). In many cases because of the complexity of the interactions, especially those across multiple scales, deterministic models are not sufficient and the optimization occurs by analytical or by a hybrid analytical-computational method. In the context of DOE, interpretation of the models, model parameters, and the effects of each of the components on the properties of the mixture is the objective of screening designs. Beyond screening designs, the prediction of the optimum from the property models can help to focus future experimentation and improve the interpretation of the property models. Reaching the optimum is more efficient if the obtained model is adequate (Lazic 2004). As discussed in Section 7.2, the F-test is most often used as a measure of the adequacy or lack-of-fit of the model. Increasing the number of model parameters is usually the preferred choice to improve the lack-of-fit, often resulting in quadratic, specialcubic, or even special-quartic models (Brandvik 1998, Brandvik and Daling 1998). For these higher order models, transformations would need to be applied in order to sufficiently linearize them for use in the property clustering algorithm. Once the model is deemed adequate, the region of applicability needs to be determined. Since the mixture design is performed in the local domain, then the resulting regression coefficients of the property models express factor effects exactly in that part of the domain. If the predicted optimum lies outside of the experimental region studied, then the experiments need to be recentered around the optimum. Many techniques to achieve this

repositioning, including gradient, nongradient, and other optimization methods, were discussed in Chapter 3. Gradient methods are based on the derivative of the response surface model and, as such, are used only when the model is deemed to be adequate. The most common gradient method is the method of steepest ascent which performs experiments at predetermined steps along the vector formed by the gradient of the response surface and is discussed in Chapter 3. The two shortcomings of the method, namely weighting fractions and visual representation, can be handled by property clustering. Since property clustering converts the properties into conserved surrogate property clusters, the steepest ascent gradient method can be entirely performed in the clustering domain at reduced complexity. The terms used in the gradient expression are the regressors corresponding to each effect direction and remain unchanged as the solution is marched toward the predicted optimum represented as the feasibility region. Furthermore, the step size can be estimated in the property domain rather than the component domain, which may decrease (or increase) the number of steps needed to reach the feasibility region. For this reason, it is beneficial to estimate both step sizes, choosing the largest step whose predicted property cluster remains within the MFR. Once the solution leaves the MFR, the method is stopped and a new MDOE will need to be conducted.

In cases where the gradient method is unable to reach the feasibility region, either the adequacy of the model needs to be improved or a more realistic set of targets needs to be specified. If the adequacy of the model cannot be improved enough to be deemed sufficient and, as such, is deemed inadequate, then nongradient optimization is performed. Nongradient optimization searches for the optimum using a step-by-step comparison of obtained property values. One of the most common types of nongradient optimizations is the simplex self-directing method (Lazic 2004). This method works by first conducting a simple mixture design around an

initial guess. Next the lowest response property value is dropped and a mirror-image experiment opposite of the dropped value is conducted. Within the property clustering construct, the mirror image may be calculated in the property cluster domain and then back calculated to get the mixture fractions. Like the gradient method, this procedure has the potential to provide larger steps toward the solution and continues until the algorithm repeats the same experiments, indicating that the optimum is somewhere in the bounded region. However, since, by definition, the use of the nongradient method is indicative of an inadequate model, then the terms of the model will not adequately convert the measured properties to equivalent constituent fractions. As such, it is this author's recommendation that property clustering should only be used as a visualization tool. Doing so allows the march toward a candidate solution for systems with multiple chemical constituents and/or multiple properties to be monitored regardless of the number of chemical constituents studied for up to three properties. For more than three properties, the algebraic method developed by Qin *et al.* (2004) can be adapted and used.

Once the boundary of the MFR has been reached, a new MDOE can be performed. Good properties of the MDOE consist of orthogonality, rotatability, and symmetry about an experimental center point (Lazic 2004). The spacing of the experimental design points around the target optimum in the new MDOE is performed in a manner similar to the original MDOE. One exception, however, is the choice of defining the experimental design region in terms of the chemical constituents or properties. Since the success of the design is judged by the ability to arrive at the feasibility region, which is defined in terms of properties, then it would follow that the new experimental design region should be conducted using properties of the feasibility region mapped back to the component space. Using these mixture ratios will most likely result in a loss of symmetry about the experimental design center point. Various optimality algorithms

describing the variance profiles associated with each type of design can be used to quantify the effect of the loss of symmetry. It is the experimenter's discretion whether to use the design points that better describe the feasibility region or the design points that are more symmetric. Once the experiments are conducted, the aCAM^bD procedure is repeated.

7.4. Case Study: Polymer Spun Yarn Mixture Design

The case study used to highlight this methodology is the development of a polymer blend of spun yarn for marine applications. The yarn is blended in a filament winding operation under the constraints of classical lamination theory, which states that the properties of the wound filament are universal throughout the yarn (Gurdal *et al* 1999). Three properties (e.g. $Y = P$ and $p = 3$) were evaluated by researchers: thread elongation (P_1), knot-strength (P_2), and density (P_3). Four components (e.g. $u = 4$) were evaluated using a simplex-lattice design of $n = 10$ experiments. Data for components $i = 1, 2, 3$ (low-density polyethylene (LDPE), polystyrene, and polypropylene) were compiled from Cornell (2002). Data for a fourth component $i = 4$, nylon 6,6, was collected from Rodriguez *et al.* (2003) The resulting MDOE is shown in Table 7.4.

Table 7.4: Simplex-Lattice Design for the MDOE of Polymer Spun Yarn

Exp. Run	Chemical Constituents				Response Properties		
	x_1	x_2	x_3	x_4	y_1	y_2	y_3
1	0.96	0.02	0.02	0.00	11.75	9.70	1.29
2	0.50	0.48	0.02	0.00	10.69	11.20	1.14
3	0.50	0.02	0.48	0.00	13.91	10.80	1.18

4	0.73	0.25	0.02	0.00	11.22	10.45	1.22
5	0.50	0.25	0.25	0.00	12.30	11.00	1.16
6	0.73	0.02	0.25	0.00	12.83	10.25	1.24
7	0.50	0.25	0.00	0.25	10.82	10.66	1.20
8	0.50	0.00	0.25	0.25	12.57	10.44	1.22
9	0.25	0.25	0.25	0.25	11.99	11.26	1.14
10	0.57	0.17	0.17	0.08	12.01	10.64	1.20

Using linear regression, first-order Scheffe and Cox models were developed for the properties thread elongation and knot strength. For the third property, density, a previously developed pure component property operator model was utilized (Eden *et al* 2003). Fitting the models to the data set in Table 7.4 resulted in significant models with all individual component terms also significant. The regression coefficients for the two model types are found in Table 7.5.

Table 7.5: Scheffe and Cox Model Regression Coefficients

<i>i</i>	Scheffe Models			Cox Models		
	$\Psi_1(\beta_i^*)$	$\Psi_2(\beta_i^*)$	$\Psi_3(\beta_i^*)$	$\Psi_1(\beta_i)$	$\Psi_2(\beta_i)$	$\Psi_3(\beta_i)$
1	11.70	9.59	1.30	-0.2968	-1.041	0.104
2	9.40	12.85	0.98	-2.597	2.219	-0.216
3	16.40	11.98	1.07	4.403	1.349	-0.216
4	10.47	10.61	1.20	-1.524	-0.02433	0.00365
s	-	-	-	12.00	10.63	1.20

To determine the specific product targets, it is important to understand the relationship between the end-use attributes and the measured properties. The product in this design is to be used in the spinnaker sheets and guys on a race sailboat. One of the important properties in this application is that the sheets and guys have some stretch so that the spinnaker will stay filled in a gust of wind, but not so much that it loses its designed aerodynamic shape. It has been determined that this attribute is best observed with a thread elongation between 12 and 16 kg of force. The sheets and guys are also under an immense load and need a high breaking strength while maintaining some flexibility. This attribute has been determined to be best represented by a knot strength between 12 and 13 lb of force. Finally, during the setting and dousing of this spinnaker on a race boat, the sheets and guys may contact the water surface. If they are too dense, they may ride under the boat and foul the keel. If they are not dense enough, their diameter could change too much when put under load. On the basis of these attributes, the specific volume should be between 1.0 and 1.25 mL/g.

Utilizing the regression coefficients found in Table 7.5 in the models of Eq. 3.3 and Eq. 3.14 results in the response surfaces in the simplex diagrams of Figure 7.11. The experimental design points and their resulting property values are also plotted. The feasibility region is the region shown in green and yellow. Immediately obvious is the difficulty in the interpretation of the overall design because of the use of multiple charts. Likewise the influence of single chemical constituents is difficult to measure because they may have competing effects for different properties; adversely effecting one property to the benefit of another. Furthermore, should an additive be chosen to supplement the design, additional figures would be required to determine its impact and relationship to the current design. All of these conditions make for a less than ideal situation known as combinatorial explosion.

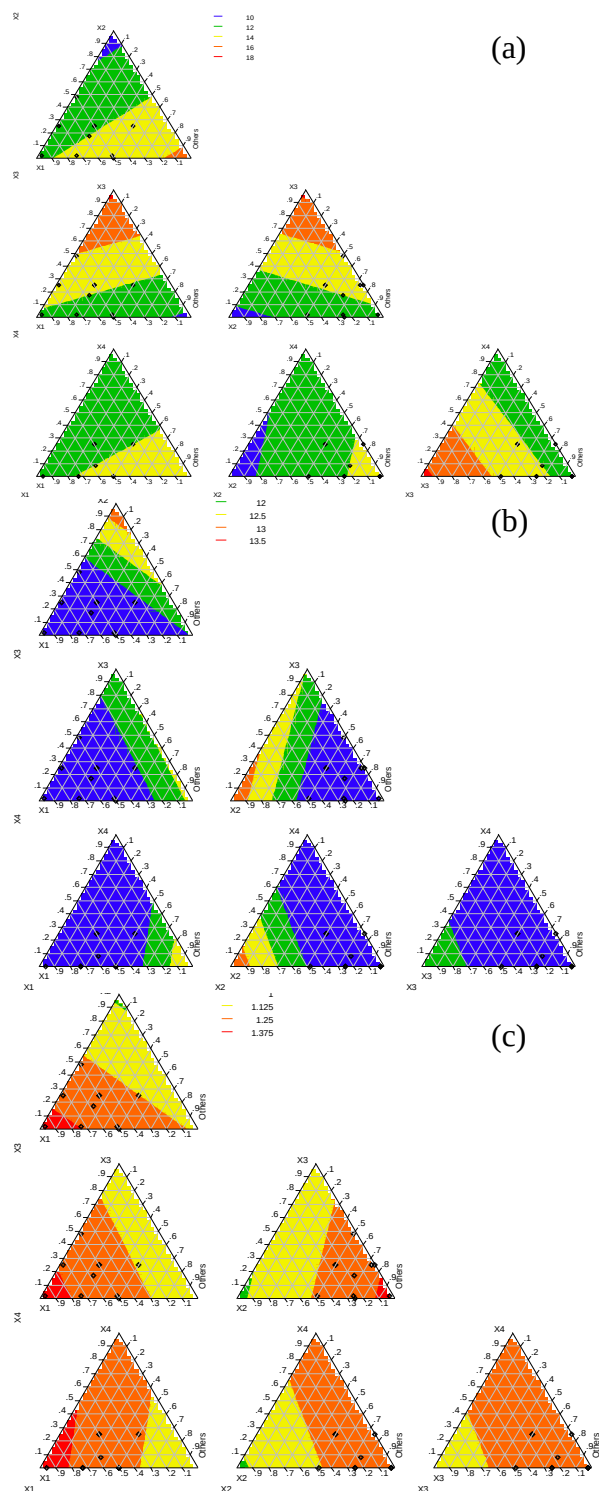


Figure 7.11: The Various Simplex Diagrams based on First Order Models of the Responses (a) Thread Elongation, (b) Knot Strength, and (c) Specific Volume. The Product Target Regions are in Green and Yellow.

To provide an easier method of which to examine the impact of components on all the properties simultaneously, the design is analyzed using property clustering. Using Eq. 7.7 for the Scheffe model and Eq. 7.8 and Eq. 7.11 for the Cox model, a set of dimensionless property operators are created with a set of references chosen to ensure a positive *AUP*, as shown in Table 7.6.

Table 7.6: Nondimensionalized Property Operators and References

<i>I</i>	Scheffe Models			Cox Models		
	Ω_1	Ω_2	Ω_3	Ω_1^z	Ω_2^z	Ω_3^z
1	0.780	0.639	1.300	-0.00278	-1.176	1.19
2	0.627	0.857	0.980	-0.0243	2.508	-2.48
3	1.093	0.799	1.070	0.0413	1.525	-1.45
4	0.698	0.707	1.200	-0.0143	-0.02749	0.0419
s	-	-	-	0.112	12.01	13.7
ref	15	15	1	107	0.8849	0.0872

The Scheffe dimensionless property operator models are then converted to clusters using Eq. 6.6 and shown in Table 7.7. Next, the responses measured at each of the mixture design points are converted to property clusters and shown in Table 7.8. Finally, the specified product targets are converted to a target feasibility region using the method discussed in Chapter 6, resulting in Table 7.9.

Table 7.7: Property Clusters and *AUPs* of the Scheffe Canonical Property Operators

<i>i</i>	Property Clusters			<i>AUP</i>
	C_1	C_2	C_3	
1	0.287	0.235	0.478	2.72
2	0.254	0.348	0.398	2.46
3	0.369	0.270	0.361	2.96
4	0.268	0.271	0.461	2.61

Table 7.8: Property Clusters and *AUPs* of the Experimental Mixtures

<i>i</i>	Property Clusters			<i>AUP</i>
	C_1	C_2	C_3	
1	0.288	0.238	0.474	2.72
2	0.274	0.287	0.439	2.60
3	0.328	0.254	0.418	2.83
4	0.281	0.262	0.457	2.66
5	0.302	0.270	0.428	2.72
6	0.308	0.246	0.445	2.77
7	0.275	0.271	0.455	2.63
8	0.305	0.253	0.442	2.75
9	0.298	0.279	0.423	2.69
10	0.296	0.262	0.442	2.71

Table 7.9: Property Clusters and AUP_F of the Feasibility Region

<i>Feasibility Region</i>	Property Clusters			AUP_F
	C_1	C_2	C_3	
Pt. 1	0.281	0.281	0.439	2.85
Pt. 2	0.274	0.297	0.429	2.92
Pt. 3	0.300	0.325	0.375	2.67
Pt. 4	0.364	0.295	0.341	2.93
Pt. 5	0.372	0.279	0.349	2.87
Pt. 6	0.342	0.257	0.401	3.12

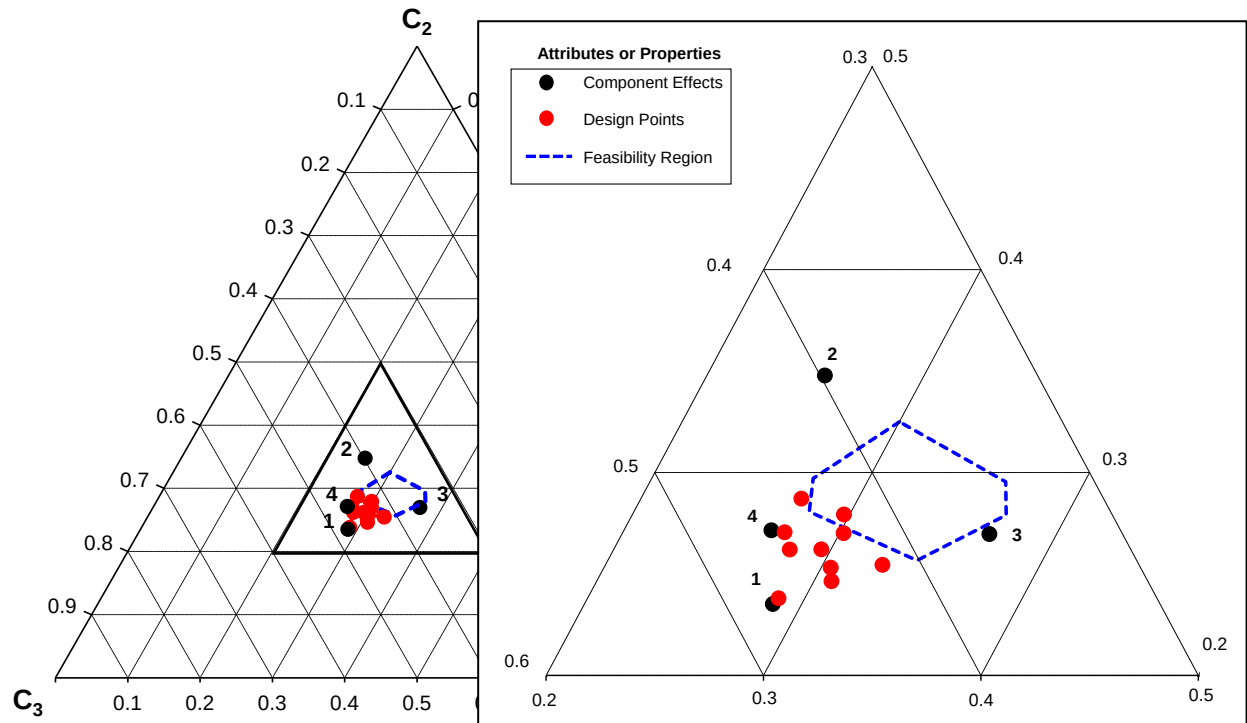


Figure 7.12: The Experimental Design Points of the Case Study, a Four-Component Mixture of Polyethylene, Polystyrene, Polypropylene, and Nylon 6,6 Mapped to Property Cluster Space (Solvason *et al.* 2008; Solvason *et al.* 2009).

All clusters are then plotted in property clustering diagrams shown in Fig. 7.12 with the vertices representing each of the three properties in their cluster forms. By reducing the complexity of the visualization of the design problem using property clustering, it is now clear that the third chemical constituent, polypropylene, is closest to the feasibility region, followed by constituent 4 (nylon 6,6), constituent 2 (polystyrene), and constituent 1 (LDPE). Thus, the addition of polypropylene to a hypothetical mixture will do little to change the properties of the mixture, suggesting it should be used as a filler. Of the three remaining polymers, LDPE appears to have the largest impact on the mixture properties when looking at Figure 7.12. However, since the component effects were derived using Scheffe models, inherent colinearities exist.

Converting from the Scheffe canonical models to Cox polynomial models with a standard reference mixture at location (0.574, 0.173, 0.173, 0.083) removes the primary colinearity resulting in a better visualization of the effect of component i .

The Cox property operators in Table 7.6 are then used to calculate the pseudo cluster with Eq. 7.16 shown in Table 7.11 and the standard cluster with Eq. 7.17 shown in Table 7.10.

Table 7.10: Property Clusters and AUP s of the Standard Reference Property Operators

i	Property Clusters			AUP
	C_1^s	C_2^s	C_3^s	
s	0.00435	0.465	0.531	25.85

Table 7.11: Property Clusters and AUP s of the Cox Polynomial Property Operators

i	Property Clusters			AUP
	C_1^z	C_2^z	C_3^z	

1	-0.2781	-117.6	118.9	0.0100
2	-18.50	1907	-1888	0.00131
3	0.3549	13.11	-12.47	0.0116
4	-142.8	-275.0	418.9	0.000100

Likewise, the pseudo feasibility region is also calculated by correcting the points listed in Table 7.9 to get Table 7.12.

Table 7.12: Property Clusters and $AUPs$ of Pseudo Feasibility Region

<i>Pseudo Feas. Region</i>	Property Clusters			AUP_F
	C_1^z	C_2^z	C_3^z	
Pt. 1	0.0040	0.4841	0.5119	28.01
Pt. 2	0.0039	0.5041	0.4921	29.14
Pt. 3	0.0043	0.5591	0.4366	26.28
Pt. 4	0.0057	0.5583	0.4360	26.31
Pt. 5	0.0060	0.5385	0.4556	25.18
Pt. 6	0.0043	0.4834	0.5112	28.05

Figure 7.13 illustrates the standard and pseudo clusters, zoomed out to include negative cluster space. The values located along the horizontal x-axis and vertical y-axis are the Cartesian coordinate scales. Here it is confirmed that polypropylene (3^z) has the smallest effect on the combined mixture properties. However, by removing most of the colinearity in the model, the result now clearly shows that polystyrene (2^z) has the strongest effect on the combined mixture

properties. It also shows that polystyrene (2^z) and LDPE (1^z) have inverse effects, a result not clear in Fig. 7.11.

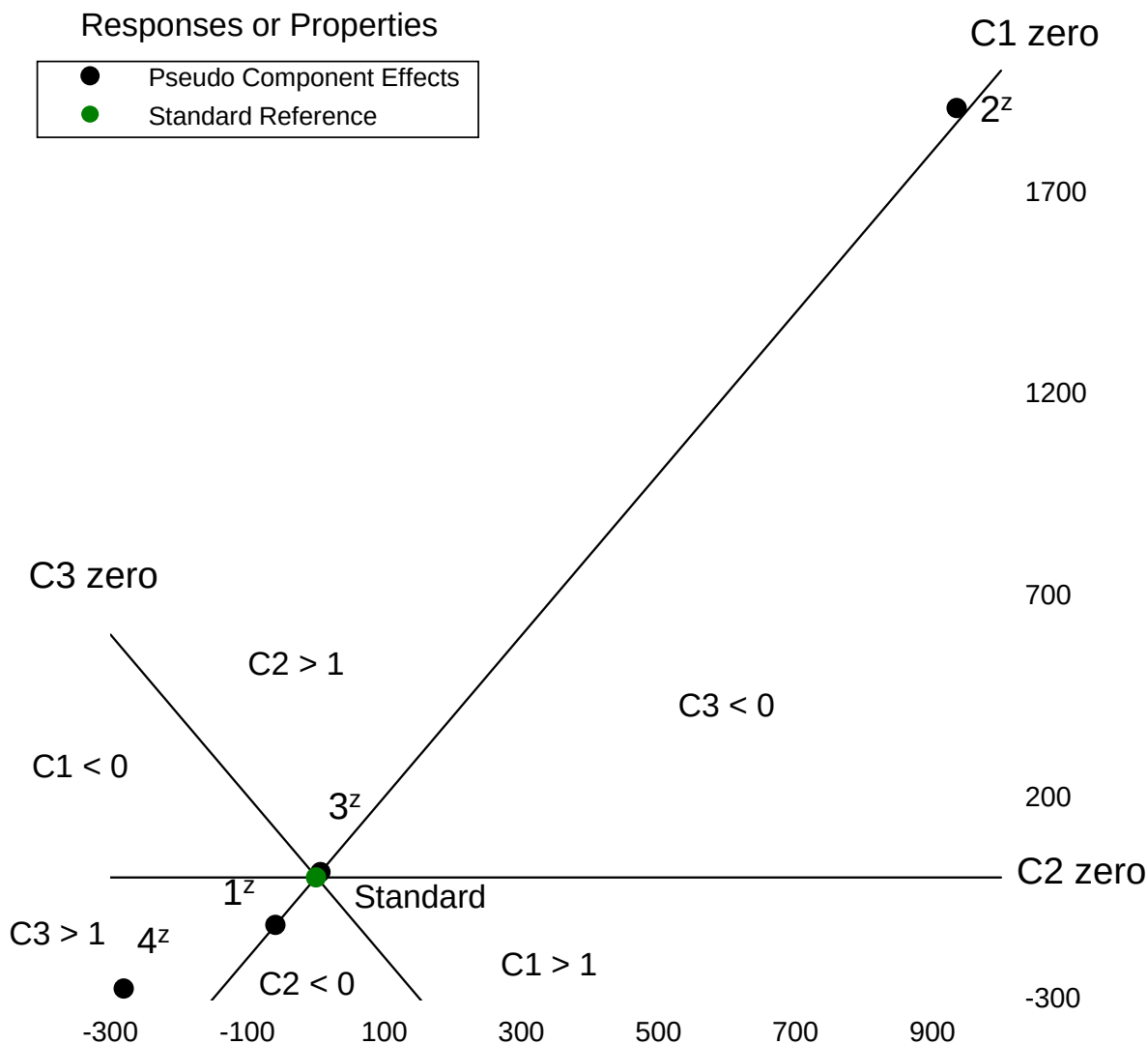


Figure 7.13: The Clustering Diagram for the Case Study Showing the Cox Polynomial Model in Negative Cluster Space.

Evaluating the placement of the experimental design points in the property cluster space also offers insights into the design. In Fig. 7.12 the experimental design points are translated to the property cluster space. A single experiment consisting of an equimolar ternary mixture falls within the feasibility region. Unfortunately, the rest of the design points are outside the

feasibility region and none of the candidate mixtures fall within the AUP range. This inference can also be made when investigating Fig 7.11 but with considerably more effort.

To find potential candidate solutions to the mixture design, the aCAM^bD was executed in accordance with Fig. 7.7 and Table 7.3. Since the calculation of canonical component clusters, feasibility clusters and AUP values required fewer steps to compute, this data set was chosen as the inputs for the aCAM^bD algorithm. Alternatively, pseudo component clusters, pseudo feasibility clusters, and AUP^Z could have also been chosen.

First, a search for pure component candidate solutions was initiated, resulting in no matches. Second, binary candidate solutions were generated. Mixtures 1-2, 1-4, and 2-4 were found to fail Rule 1, so they were excluded prior to property calculation. Mixtures of components 1-3, 2-3, and 3-4 were found to meet Rule 1 and Rule 2. However, when mixture 1-3 was tested against Rule 3, it was discovered that its mixture AUP_M range did not overlap the feasibility region's AUP_F along its mixing curve and was thus discarded. The AUP_M range of mixture 3-4 also failed to overlap the feasibility region's AUP_F range along its mixing curve and was likewise discarded. Fortunately, the AUP_M range of mixture 2-3 did overlap the feasibility region's AUP range, and was thus back calculated by setting the AUP_M to the AUP_F of the feasibility region at the intersection point. It was determined that the mixture range varied primarily in C_1 along C_2^U and C_3^L , so the corresponding mixture was determined using the AUP_F of Pt. 3 and Pt. 4, respectively. The resulting mixture fractions were calculated with Eq. 6.33 to get $0.057 < x_2 < 0.591$ and $0.409 < x_3 < 0.943$, as shown in Table 7.13. Mixture 2-3 represents the simplest mixture that delivers the desired attributes of the design problem.

Table 7.13: Binary Mixtures that met Rules 1 and 2.

Binary Mixture <i>i-w</i>	<i>AUP_M</i>		<i>AUP_F</i>		Mass Fraction <i>i</i>	
	<i>LL</i>	<i>UL</i>	<i>LL</i>	<i>UL</i>	<i>LL</i>	<i>UL</i>
1-3	2.87	2.89	3.07	3.11	-	-
2-3	2.62	2.94	2.71	2.93	0.057	0.592
3-4	2.73	2.93	2.96	2.97	-	-

Any binary mixture candidate will also be able to form a ternary mixture candidate. Therefore, ternary mixtures 1-2-3 and 2-3-4 will, by definition, meet Rules 1, 2 and 3. To determine the maximum amount of component 1 that can be added to the solution, the AUP_F was first calculated by looking for a minimum Ω_2 along the mixing lines between component 1 and the 2-3 binary mixture. The resulting mixtures produced maximums of $x_1 \leq 0.012$ and $x_1 \leq 0.168$, with the largest fraction being returned. The procedure was repeated for mixture 2-3-4 and the resulting mass fractions are shown in Table. 7.14.

Table 7.14: Ternary Mixtures that met Rule 1, 2, and 3.

Ternary Mixture <i>i-w-r</i>	Mass Fraction <i>i</i>		Mass Fraction <i>w</i>	
	<i>LL</i>	<i>UL</i>	<i>LL</i>	<i>UL</i>
1-2-3	0.00	0.168	0.048	0.592
2-3-4	0.056	0.592	0.400	0.943

It should be noted that the procedure to find the AUP_F is cumbersome. An often quicker alternative is to predict the property values at the intersection point of the feasibility region and

the mixture curve and march the solution along the curve until one of the properties no longer falls within the feasibility limits. This point occurs at the location where AUP_F is equal to AUP_M .

In addition to the ternary mixtures shown in Table 7.13, any binary mixture that met Rule 1, but did not meet Rules 2 or 3, may also be able to meet the constraints on the design. For example, the ternary mixture 1-3-4 is comprised of two binary mixtures (1-3 and 3-4) that met Rules 1 and 2, but failed to meet Rule 3. Selecting the binary mixture with the broadest mass fraction delta that met Rule 2 (e.g. mixture 3-4) and investigating the remaining pure component (e.g. component 1) results in an indeterminate AUP_F . This result is directly attributable to all Ω_2 values used in the mixture being below the targets for the design. As a result, mixture 1-3-4 is not a candidate solution. The procedure can also be applied to a tertiary mixture, but was beyond the scope of this design.

A model feasibility region (MFR) can also be calculated for this aCAM^bD. In particular, the $a = 3$ and $n = 10$ multivariate system, a MFR was estimated at 95% confidence around each of the design points as shown in Fig. 7.9. As shown in the figure, the boundary of the closest candidate solution, a ternary mixture of component 1, 2, and 3 still remains outside of the MFR, failing to meet Rule 6, meaning that the property operator models must be partially extrapolated to estimate the candidate. This is an insufficient design. To prevent model extrapolation, the design points should be repositioned so that they cover the entire feasibility region. The procedure for executing the repositioning must take into consideration the increase in accuracy of the property space at the expense of optimality of the component space. Two methods exist for the reparameterization: (1) increase the complexity of the model to improve its fit or (2) improve the estimates of the component effects by repositioning the effects over the optimum.

Since the fit of the first-order models were found to be excellent with a ξ -value of <0.0001 , then the model structure is adequate and should not be changed. Instead, new estimates of the component effects should be found by conducting a new MDOE. The center point of the new MDOE should be at the optimum of design.

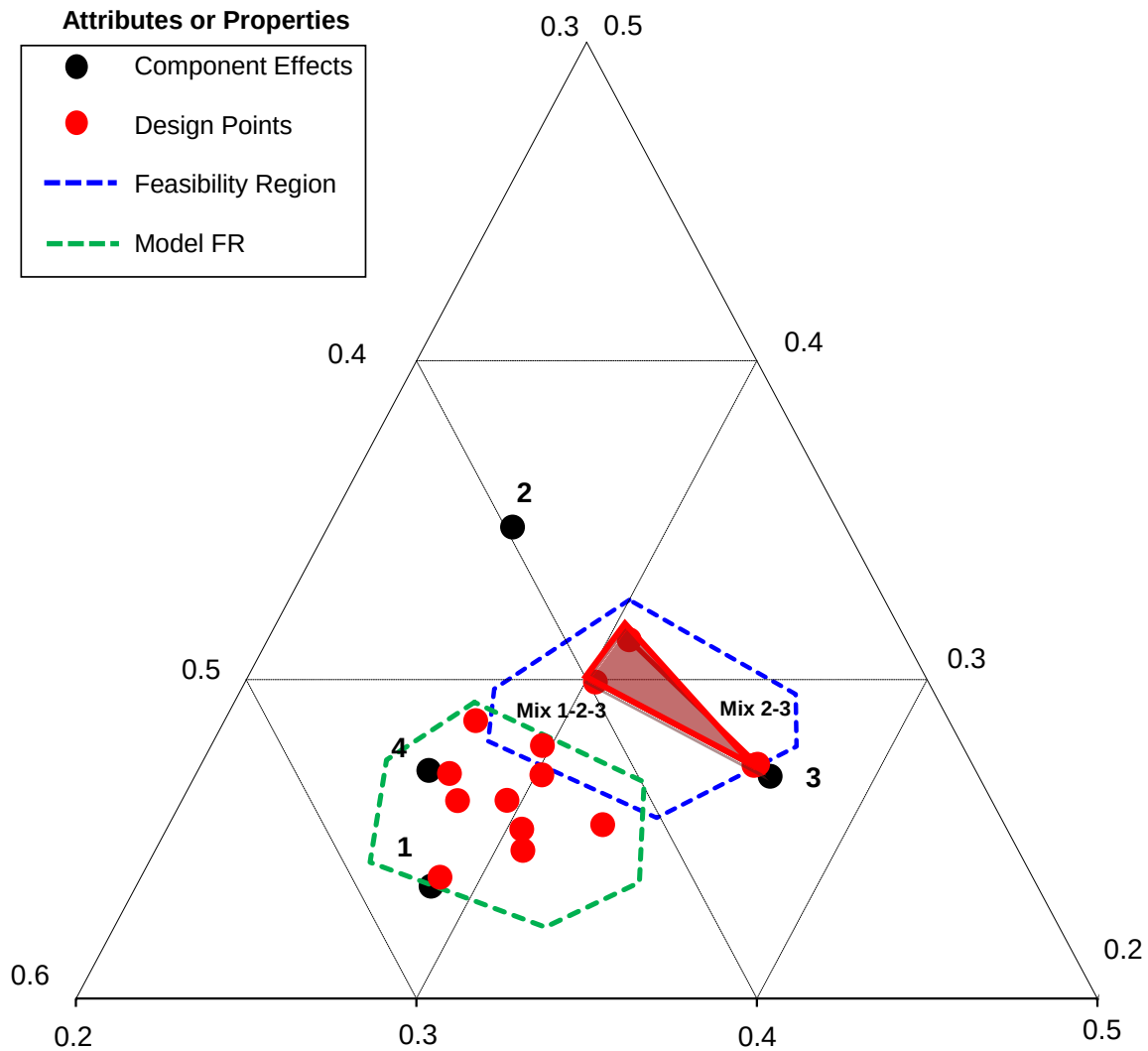


Figure 7.14: The Model Feasibility Region (MFR) and Candidate Solutions.

It can be found in two ways: (1) using the existing response surface models and (2) performing a domain search using a gradient method. For the first method, the Solver function in Excel was used to find the optimum by maximizing a geometric mean square of the system of response models. The optimization resulted in an optimum located at chemical constituent 3, polypropylene. The magnitude of the factor level adjustment for the MDOE is then at the discretion of the experimenter. For the second method, the gradient direction of the combined response surface was found to be equal to the values regressors, which suggests a vector pointing toward polypropylene. The choice of step size along the gradient is again at the discretion of the experimenter. Both of these methods require additional experiments that are beyond the ability of our laboratory to perform and as such limits further analysis of the case study. However, the main objective regarding the use of property clustering to reduce combinatorial explosion in screening designs has been thoroughly investigated. Continued analysis on optimization, decomposition methods, and phenomenological models is presented in Chapter 8 and Chapter 9. Regardless of the applicability of the response surface models, clustering the solution still provides the benefit of viewing the design in its entirety on a single diagram, irrespective of the number of components studied as long as the number of properties measured is three or less. In cases when three or more properties are studied, additional diagrams may be used or algebraic methods applied.

7.5. Summary

In this work, a systematic property-based framework for evaluation of mixture design problems using property clustering has been presented. The recently introduced property integration framework has been extended to include experimentally derived property operator models: specifically first-order Scheffe canonical and Cox polynomial models. When

interpretation of the chemical constituent's impact on the mixture property is warranted, Cox derived property operator models are utilized such that the location of the pseudo chemical constituent relative to the standard reference mixture is indicative of its impact on the mixture's properties. The accuracy of the design is visually observed by placing the design points in the property design space.

A significant result of the developed methodology is that for problems that can be satisfactorily described by just three properties, the experimental mixture design problems are analyzed visually on a simplex diagram, irrespective of how many chemical constituents are included in the search space. However, algebraic- and optimization-based approaches can easily extend the application range to include more properties. In summary, the technique proposed should be capable of extending the use of the property clustering algorithm and the reverse problem formulation to include the utilization of data and data-driven models. This significant contribution will allow the exploration of models capable of describing properties that depend on molecular architectures that may exist at a variety of scales, so long as they can be linearized. In particular, difficult to quantify chemical product attributes that rely on experimental descriptions, such as toxicity or disintegration time, can now be simultaneously evaluated in a property cluster diagram, providing valuable design insights. Furthermore, the use of data driven models opens up the clustering framework to instant validation using known property values of chemical products. However, the method outlined will be limited to pure components and mixtures of pure components, which will limit the range of chemical products that can be designed. In order to extend the design range, smaller building blocks that could be combined to build molecules not present in the experimental design would be useful. One such technique is the group contribution method (GCM) which is a QSPR technique that distills chemical property data

down to individual functional groups using group theory. Combining data-driven techniques with GCM within the clustering framework and applying it to a reverse problem formulation to investigate all chemical product architectures within a specific property range will be discussed in detail in Chapter 8.

Chapter 8

Attribute-Computer Aided Molecular Design (aCAMD)

In his review of chemical product design, Gani (2004) defines the general structure of product design problems in process systems engineering, identifying several areas that the systems engineering community can improve through the development of novel methods and tools. One particular need identified by Gani (2004) was for a method to enlarge the application range of existing property models and/or to develop new property models in order to better describe structured chemical products with large numbers of atoms or highly electronegative behavior. In the absence of adequate models, the conventional approach to these types of problems has been to rely on empirical or simulation models to describe the system, estimating the boundaries where the models are adequate, and then designing the molecular architecture of chemical products in those adequate domains. While accurate, this approach is ill-formed, time-consuming, expensive, and overly narrows the design parameters, resulting in non-optimum products. As shown by Duvedi and Achenie (1996), the application of conventional solvers like mixed-integer, non-linear programs (MINLPs) to a computer aided molecular design (CAMD) is limited to local optima solutions when the property prediction methods are non-linear and there is no guarantee that the global optimum exists among the generated list of candidate molecules. Although work in disjunctive programming (Grossman 1999, Sammons *et al.* 2009, Odjo *et al.* 2011) has been developed to address some of these shortcomings, the uncertainty in the design

space, lack of insight in to the relative impacts, similiarities, and interactions of each of the candidate molecules and molecular groups, and forward problem orientation limit its use.

Recognizing the above limitations, this chapter describes a new technique called attribute based computer aided molecular design (aCAMD) that combines experimental methods with CAMD techniques in a reverse problem formulation (RPF) to obtain the enumeration of all possible candidates and candidate mixtures. In Section 8.1, the mapping of data in the consumer attribute domain in to a physical-chemical property domain described by group contribution methods (GCM) is discussed followed by the identification of suitable alternatives using 1st and 2nd order GCM techniques in the property clustering framework. Section 8.2 discusses the clustering of the pure component effects, candidate mixtures, feasibility region, and the model feasibility region. Section 8.3 describes the aCAMD solution approach, interpretation of the generated candidates, the model feasibility region (MFR), and the limitations of the technique when cross-validating the results in both the property and attribute domains. Section 8.4 highlights the method in a case study on environmentally friendly refrigerant design and Section 8.5 concludes the chapter with a summary.

8.1. Integrating Attribute Data and GCM Models with the RPF

It is well recognized that experimental methods yield probabilistic models which may be of less accuracy than computational chemistry based deterministic models. However, several well established techniques exist for the proper estimation of this residual error and its resulting propagation. In addition the probabilistic models can be tailored to have a reduced non-linearity which results in fewer local optima and improve the likelihood of finding the global optima. If any non-linearities exist in the constitutive equations, they can be marginalized by using a reverse problem formulation with property clusters to solve the problem in the property design

space by making use of the duality of linear programming (Eden *et al.* 2004, Eljack *et al.* 2007). Enumeration of all possible candidate mixtures is possible using group contribution methods (GCM) to predict the molecular structure and by matching the result against the product design sub-problem. This allows for the complete identification of all possible candidates and candidate mixtures without being computationally expensive. The benefit of using GCM is that the additives which are not part of the original data set can be predicted, reducing the number of experiments needed to quantify the attribute-property-constituent relationships. Additionally, insights into the molecular structures of the candidates and candidate mixtures can be obtained. Finally, the method shown here demonstrates that mapping design information from one property domain to another is feasible, which is an important key to solving multi-scale chemical product design problems.

In product design the key to successfully designing a molecule that meets a set of constraints is the ability to adequately predict the molecular structures and properties. Traditionally, GCM has been used to predict the physical properties of a molecule based on the additive nature of individual group fragments. Unfortunately, the method is limited by the number of properties it is able to predict. For properties not described by GCM (i.e. not in Table 6.1), a relationship between non-GCM described properties and GCM described properties must be obtained (Eljack *et al.* 2007). Often the non-GCM described properties are consumer attributes. Consumer attributes consist of a list of product characteristics the consumer finds desirable or undesirable. The attributes are typically modeled as a list of empirical and theoretical equations of known and controllable physical properties. Represented mathematically, there is a set of attributes **A** which are functions of physical properties **P** (Solvason *et al.* 2009).

$$A = f(P) \tag{8.1}$$

The physical properties themselves can be expressed as functions of their chemical signature. These equations can be of the group contribution type (i.e. Eq. 8.2) or empirically derived (i.e. Eq. 8.3) (Joback and Reid 1987; Constantinou and Gani 1994; Marrero and Gani 2001). The expressions are often highly non-linear (Solvason *et al.* 2009).

$$P = f(n_{g_1}, n_{g_2}) \tag{8.2}$$

$$P = f(X) \tag{8.3}$$

where the properties $\mathbf{P}$ are functions of a set of chemical components $\mathbf{X}$. The property-component relationship typically contributes to the non-linear part of an MINLP. As stated earlier the solution to the MINLP is not guaranteed to be the global optimum (Duvedi and Achenie 1996). An elegant reparameterization of the MINLP is the reverse problem formulation presented by Eden *et al.* (2004) and Eljack *et al.* (2005). In this work, the process design problem is separated from the molecular design problem. The reverse solution of the process design problem generates the design targets for the molecular design problem. The molecular design problem is solved in reverse to match the design targets from the process design. The method can be extended to the attribute problem such that the solution of Eq. 8.1 determines the property targets used in the reverse solution of Eq. 8.2 and Eq. 8.3. This approach is depicted in Fig. 8.1.

For the special situations where the selection of the optimum candidate mixtures from a known data set is deemed sufficient, the MINLP and reverse problem formulation are unnecessary. In this case the attributes $\mathbf{A}$ can be directly regressed in terms of the components $\mathbf{X}$ and the component mixtures (Solvason *et al.* 2009):

$$A = f(X) \quad (8.4)$$

which is a special case of Eq. 3.5 with $Y = A$ that was discussed thoroughly in Chapter 7 and will not be discussed further in this chapter.

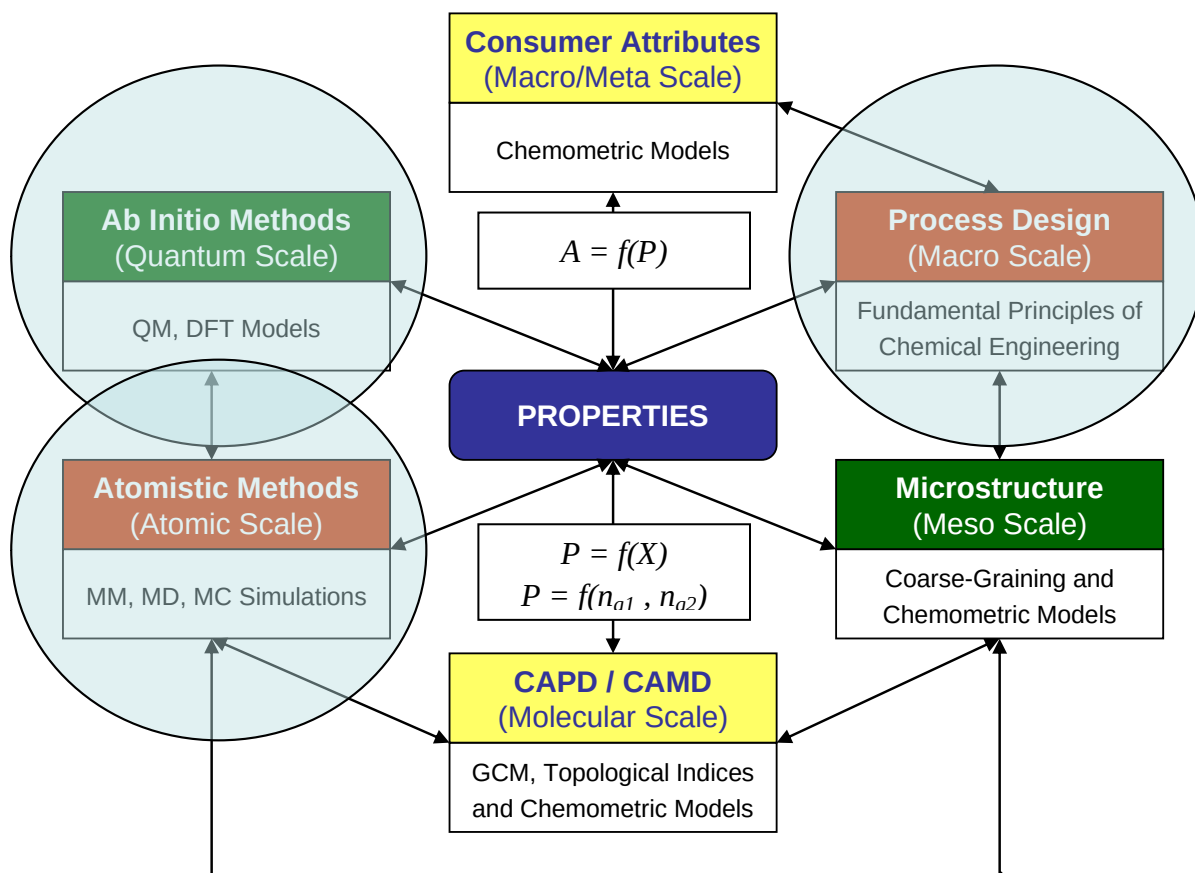


Figure 8.1: Description of the Reverse Problem Formulated Product Design with Property Mapping to Group Contribution Described Properties.

Since the objective of most molecular generation techniques is to develop a complete set of candidates and candidate mixtures, the use of property prediction techniques based on molecular structure is preferred. In this approach a set of target properties is estimated from a set of attributes subject to the following constraint (Solvason *et al.* 2009):

$$A_k = f(P_j) \text{ where } k \geq j \quad (8.5)$$

As long as the number of attributes is greater than the number of properties, unique property values can be calculated from the attribute data. Depending on the nature of the relationship either a linear program (LP) or non-linear program (NLP) must be solved such that a set of properties is estimated for each set of attributes. While a LP is preferred for its solution efficiency, most attributes are nonlinear functions of physical-chemical properties. Unfortunately, the use of a NLP introduces a potential to reach non-global solutions to the candidate generation problem. An alternative method is to use inverse regression, especially in situations where the attribute-property relationship is unknown. The use of regression to develop this relationship can reduce the complexity of the design problem at the expense of introducing uncertainty in to the solution. This uncertainty is best quantified using analysis of variance (ANOVA) and it is essential that limitations be placed on the design space such that only the well correlated regions are investigated. Similar to the attribute-based “model” feasibility region (MFR) described in Chapter 7, confidence intervals on the estimates of the attributes, AM_k^L and AM_k^U , can be found using Hotelling’s T^2 test (Stine 2001, Ramirez 2009). The aMFR upper and lower limit for the k^{th} attribute is shown in Eq. 8.6:

$$Y_{ke} \pm \sqrt{\left(\frac{a(n-u-1)}{(n-u-a)} F_{a,n-u-a}(\xi) \right) \cdot \left(\frac{n}{n-u-1} \right) \cdot S_{kk}^E \cdot x_e' (X'X)^{-1} x_e} \quad (7.27)$$

Where the response $Y = A$ for cosumer attributes. The upper and lower limits estimated by this procedure can then be used to develop an attribute “model” feasibility range (aMFR) by combining Eq. 7.27 with the responses at the design points and then selecting the maximum and minimum of the responses to get Eq. 8.6:

$$AM_k^L \leq A_k \leq AM_k^U \quad (8.6)$$

The aMFR is then created by inserting the bounds of Eq. 8.6 into Eq. 6.7. Any candidates or candidate mixtures found outside the region are known to be invalid within a $1-\xi$ confidence. Similarly, a property “model” feasibility region (pMFR) can be created by applying Eq. 7.27 and to the property responses (i.e. $Y = P$) of the design points in an experimental design, resulting in Eq. 8.7. The pMFR upper and lower limit for the j^{th} property are as follows:

$$P_{je} \pm \sqrt{\left(\frac{p(n-u-1)}{(n-u-p)} F_{p,n-u-p}(\xi)\right) \cdot \left(\frac{n}{n-u-1}\right) \cdot S_{jj}^E \cdot x_e'(XX)^{-1} x_e} \quad (8.7)$$

To determine the property target feasibility region (pFR), attribute targets set by the consumer that define an attribute feasibility region (aFR) are transformed into property targets using an inverse regression expression of Eq. 8.5 resulting in the following:

$$P = f(A) \quad (8.8)$$

where the parameters of the fitted model are written as:

$$B = (A'A)^{-1} A'P \quad (8.9)$$

Like the pMFR, the pFR confidence interval grows wider as the model moves farther away from the mean of the predictor. However, since the pFR is developed from a regression model serving a predictive role, new observations must include both the error associated with the model and error associated with the new experimental observation. This results in pFR upper and lower limits for the j^{th} property as follows:

$$P_{je} \pm \sqrt{\left(\frac{p(n-a-1)}{(n-a-p)} F_{p,n-a-p}(\xi)\right) \cdot \left(\frac{n}{n-a-1}\right) \cdot S_{jj}^E \cdot (1 + A_e'(A'A)^{-1} A_e)} \quad (8.10)$$

A set of attribute targets specified by the consumer, as shown in Eq. 8.11, can then be converted into a property domain by taking the minimum and maximum property values generated using Eq. 8.8 and 8.10.

$$A_k^L \leq A_k \leq A_k^U \quad (8.11)$$

$$P_j^L \leq P_j \leq P_j^U \quad (4.16)$$

The resulting property domain is only valid if the attributes and properties can be considered unbiased, otherwise the NLP will need to be solved in the conventional manner (van Belle 2002). The benefit of eliminating the need to solve the NLP is that a global optimum solution will always be obtained, subject to a given function of the uncertainty in the regression expression. As before, the measure of model fitness is R^2 and the predicted fit Q^2 can be used to infer the accuracy of the obtained models described by Eq. 8.8; a large R^2 represents a well fit model and a large Q^2 represents a greater likelihood that the candidate solutions generated by the aCAMD algorithm will be the true global candidate solutions (Solvason *et al.* 2009). Alternatively, Eq.8.8 may be fit univariately, in which case the variance would not be pooled, and each j property would have independent R_j^2 and Q_j^2 that could be combined using a geometric mean square.

$$R_p^2 = \prod_j^p (R_j^2)^{\frac{1}{p}} \quad (8.12)$$

$$Q_p^2 = \prod_j^p (Q_j^2)^{\frac{1}{p}} \quad (8.13)$$

Other model fitness measures that use Wilks' Lamda, Pillai's Trace, Hoetelling-Lawley Trace, Roy's Greatest Root, etc. can be utilized to to improve model fitness and prediction fitness and can be found in Johnson and Wichern (2007).

8.2. Attribute and Property Clustering Algorithms

Once the property responses to the experiments are conducted and the property targets are calculated, property clustering techniques are employed to cluster the property data as previously outlined in Chapter 6. These techniques convert the property targets into conserved surrogate clusters that are described by property operators, which have linear mixing rules, even if the operators themselves are non-linear (Eden *et al.*, 2004; Shelley and El-Halwagi, 2000). Two model types can be used to design molecules in the property domain: mixture property models and group contribution models. The choice of property model dictates the type of design being conducted, a mixture property model results in a CAM^bD and a molecular property model results in a CAMD. When the mixture property model is chosen, the pure component clusters can either be (1) calculated using the property expressions in Table 6.1 for each pure component used in the design as shown in Chapter 6 or (2) by performing regression to develop the pure component effects as shown in Chapter 7. The operators for option 1 are shown in Eq. 8.14.

$$\Omega_{ij} = \frac{\beta_{ij}^*}{\psi_j^{ref}(P_j)} \quad (8.14)$$

The operators for option 2 are shown in Eq. 8.15.

$$\Omega_{ij} = \frac{\psi_j(P_j)_i}{\psi_j^{ref}(P_j)} \quad (8.15)$$

For option 1, the procedure for finding candidate mixtures is given in Section 6.2. For option 2, the procedure is similar to that presented in Section 7.3, but with the non-dimensionalized property operators replacing the non-dimensionalized attribute operators, and the FR being replaced with a pFR which is calculated using Eq. 8.10.

Both of these mixing design methods are limited to finding candidates having the same molecular architecture commonalities as the training set, and they require new experiments to make informed predictions on potential additives which can result in combinatorial explosion. For example, suppose there is interest in finding a new refrigerant and the intention is to select the optimum candidate or candidate mixture from a list of refrigerants. To solve this design problem using mixture design, the properties must first be directly estimated from the component fractions or obtained from a database. For the case study presented in Section 8.4, it is projected that a comprehensive list would contain approximately 2050 possible candidate molecules (Solvason *et al.* 2009). An explicit model relating these candidates directly to the properties would then require a minimum of 2050 experiments to quantify the 2050 parameters in the linear mixture model unless a database or decomposition is used (Solvason *et al.* 2009). Alternatively, suppose a deterministic, equation of state model like Peng-Robinson is used to evaluate the thermodynamic properties of the refrigerants and that binary interaction parameters of the model are readily available for all of the refrigerants (Brown 2007). Even in this ideal situation, the relationship between the properties of each refrigerant and the product attributes will have to be quantified through an experimental design. However, due to the non-linear nature of the property-chemical constituent relationship, mixtures of constituents cannot be used to explicitly reduce the attribute search space without sacrificing computational efficiency. Thus, at a minimum, a mixture design including all 2050 pure components plus additional parameters for estimating the non-linearity would be necessary unless the design search space was extensively restricted. Since neither of these two results is acceptable, it is preferred to use models based on group theory.

In the group contribution methods (GCM), the property function of a compound is estimated as the summation of property contributions of all the molecular groups present in the molecular structure (Ambrose, 1978, 1980; Joback & Reid, 1983):

$$P^M = \overbrace{\sum_{g_1} n_{g_1} P_{g_1}}^{1^{\text{st}} \text{ Order}} + w_S \overbrace{\sum_{g_2} n_{g_2} P_{g_2}}^{2^{\text{nd}} \text{ Order}} + w_T \overbrace{\sum_{g_3} n_{g_3} P_{g_3}}^{3^{\text{rd}} \text{ Order}} \quad (4.15)$$

where P^M is a function of property group contribution properties P_{g_1} , P_{g_2} , and P_{g_3} of the 1st, 2nd, and 3rd order groups g_1 , g_2 , and g_3 which occurs n_{g_1} , n_{g_2} , and n_{g_3} times (Eljack *et al.* 2007; Solvason *et al.* 2009; Chemmangattuvalappil *et al.* 2010). In Chapter 6 it was shown that the definition of the property models in GCM was similar to the property operators used in the clustering framework for the 1st order group contribution. The estimate of the j normalized property operator based on first order groups is shown in Eq. 6.37 (Eljack *et al.* 2007; Chemmangattuvalappil *et al.* 2009):

$$\Omega_{ji}^M = \sum_{g_1=1}^F n_{ig_1} \Omega_{ijg_1}^M \quad (6.37)$$

where, Ω_{ijg_1} is the normalized property operator of first order group, g_1 . The resulting molecular cluster is then written as Eq. 6.38:

$$C_{ji}^M = \frac{\sum_{g_1=1}^F n_{ig_1} \cdot \Omega_{ijg_1}^M}{AUP_i^M} \quad (6.38)$$

where the molecular AUP of molecule i is written as Eq. 8.16:

$$AUP_i^M = \sum_{g_1=1}^F n_{ig_1} AUP_{ig_1}^M \quad (8.16)$$

The accuracy of this approach can be improved by including second order groups. These groups can be incorporated into the cluster space by taking advantage of their linear additive rules. In addition, since second order groups have first order groups as building blocks, then it follows that second order groups can be considered as combinations of different first order groups (Constantinou and Gani 2001). Thus, the number and type of second order groups can be directly estimated from the number and type of first order groups, subject to the following rules:

Rule 11. Second order groups have first order groups as their building blocks.

Rule 12. A second order group is formed only if it has at a minimum all of the first order groups in the required number. For instance, to form the second order group $\text{CH}(\text{CH}_3)\text{CH}(\text{CH}_3)$, there must be two $-\text{CH}-$ and two (CH_3) groups.

Rule 13. One first order group can be a part of more than one second order group. If, however, one second order group is completely overlapped by another second order group, only the contribution from larger group is considered. This is because the interaction defined by the smaller group will be taken into account by the larger group. For example, if NH_3CHOH and CHOH groups are present, only the contribution from NH_3CHOH group is considered for property estimation.

The n_{ig2} number of g_2 second order groups in molecule i can be estimated using Eq. 8.17.

$$n_{ig2} = \text{Int} \left(\text{Min} \left(\frac{n_{ig1}}{\eta_{ig1}} : \frac{n_{iF}}{\eta_{iF}} \right) \right) \quad (8.17)$$

where $n_{ig1}:n_{iF}$ are the number of each g_1 first order group found in molecule i and n_{iF} and $\eta_{ig1}:\eta_{iF}$ is the total number of occurrences of each g_1 in g_2 in molecule i . As Rule 12 suggests, the minimum of these ratios will result in the actual number of second order groups present in the

molecule (Solvason *et al.* 2009, Chemmangattuvalappil *et al.* 2010). n_{ig2} is then rounded down to the nearest integer number because the number of second order groups cannot be a fractional number. The normalized property operator for the second order property contributions Ω_{ij2} is then written as

$$\Omega_{ij2}^M = \sum_{g2=1}^G n_{ig2} \Omega_{jg2}^M \quad (8.18)$$

where Ω_{jg2} is the property contribution from the second order groups. Eq. 8.18 can predict the property contribution from second order groups in most molecules as demonstrated by Cases A and B in Fig. 8.2.

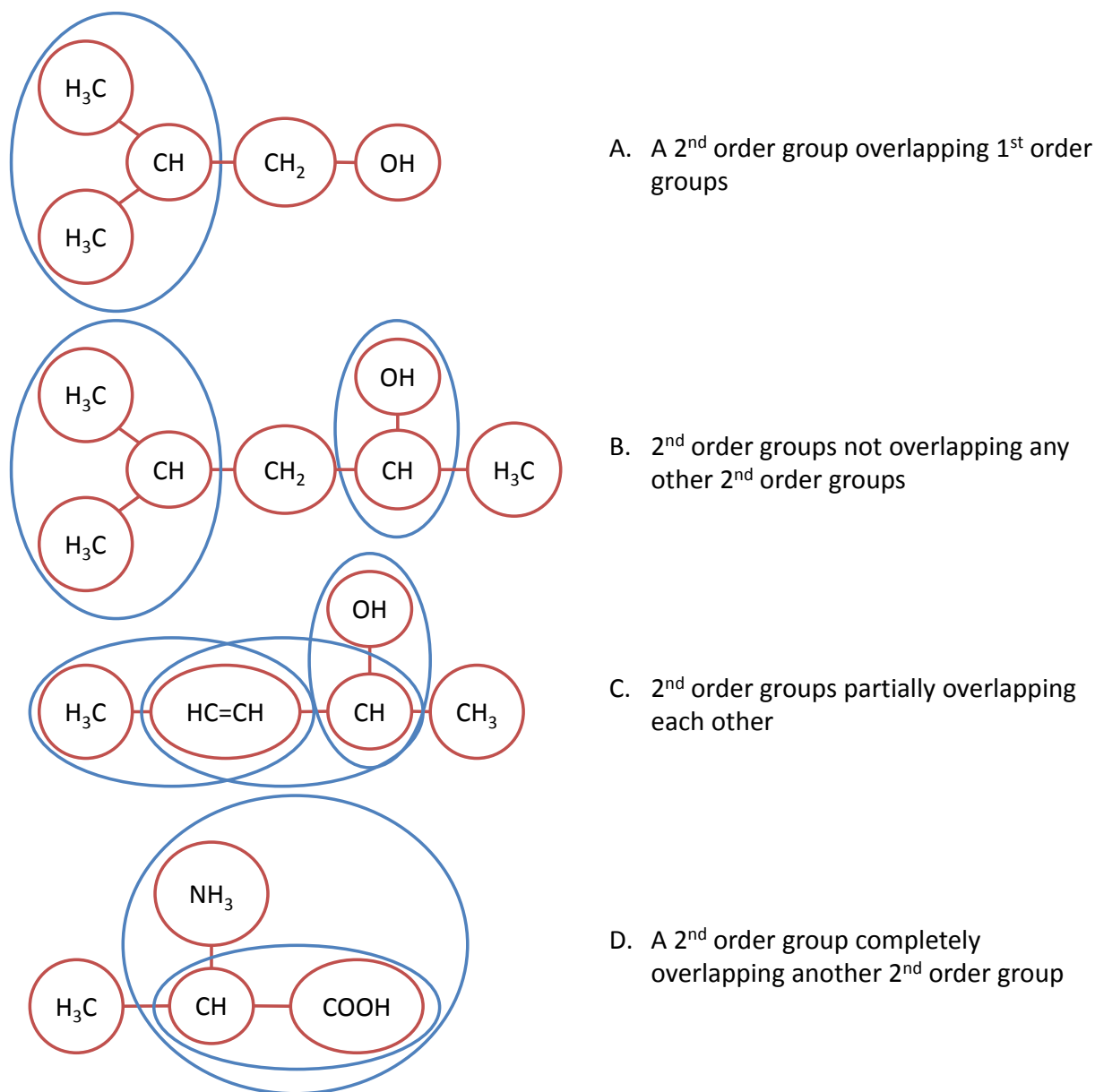


Figure 8.2: Different Situations Related to Overlapping of Second-Order Groups (Marrero and Gani 2001).

In rare occasions one or more of the second order groups will partially overlap one another (e.g. Case C) and/or completely overlap one smaller second order group (e.g. Case D) (Marrero and Gani 2001). Two methods exist for estimating Case C. The first method finds the maximum number of non-overlapped second order groups that describe the molecule, and drops the remainder. This method is discussed in more detail in Chapter 9 for IR/NIR groups.

Although valid, it has been shown that using the contributions from all groups gives better accuracy, since they generally provide dissimilar information about the molecular structure (Marrero and Gani 2001). When one molecular group completely overlaps another (i.e. Case D), Rule 13 states that the contribution of the largest group must be chosen and the contribution from the smaller group must be removed. However, it is important not to remove the group entirely from the molecule as it may appear in other locations outside of the completely overlapped group. Algebraically, this can be accomplished by removing the group from the set of G in Eq. 8.18 and then adding back only the unoverlapped number of that group using Eq. 8.19 and Eq. 8.20:

$$n_{ig2}^* = \text{Int} \left[\text{Min} \left(\frac{n_{is2}}{\eta_{g_2v_2}} : \frac{n_{iG}}{\eta_{g_2G}} \right) - \text{Min} \left(\frac{n_{ig1}}{\eta_{ig1}} : \frac{n_{iF}}{\eta_{iF}} \right) \right] \quad (8.19)$$

where $n_{iv2}:n_{iG}$ is the number of each dissimilar v_2 second order group in molecule i and $n_{g2v2}:n_{g2G}$ is the number of each dissimilar v_2 second order group completely overlapped by second order group g_2 (Chemangattuvalappil *et al.* 2009). If the smaller second order group only appears in the overlapped group, the two terms in the right hand side of Eq. 8.19 will cancel out. Otherwise, Eq. 8.19 estimates the number of occurrences of each group remaining in the molecule which are not completely overlapped. If Ω_{jg2}^* is the contribution from the smaller overlapped second order groups, then the normalized property operator for the property contributions from smaller second order groups, Ω_{ij2}^* can be calculated as the following (Solvason *et al.* 2009, Chemangattuvalappil *et al.* 2010):

$$\Omega_{ij2}^{M*} = \sum_{g^2=1}^{G_U} n_{ig2}^* \Omega_{jg2}^{M*} \quad (8.20)$$

where $g_2:G_U$ are the set of non-overlapped second order groups and $g_2:G_O$ are the set of overlapped second order groups.

$$G = G_U + G_O \quad (8.21)$$

Eq. 8.18 is then written as the following:

$$\Omega_{ij2}^M = \sum_{g2=1}^{G_O} n_{ig2} \Omega_{jg2}^M \quad (8.22)$$

The normalized property operator for molecule i can now be estimated as (Solvason *et al.* 2009; Chemmangattuvalappil *et al.* 2010):

$$\Omega_{ij}^M = \Omega_{ij1}^M + \Omega_{ij2}^M + \Omega_{ij2}^{M*} \quad (8.23)$$

Thus, given a set of molecular groups, candidate chemical structures can be estimated from first order group contributions, second order group contributions containing overlaps, and contributions from non-overlapped occurrences of second order group contributions overlapped by larger second order overlaps. Clustering the non-dimensionalized property operators of Eq. 8.23 using Eq. 6.5 and 6.6 results in the clusters shown in Eq. 8.24-8.26.

$$C_{ij1}^M = \frac{\Omega_{ij1}^M}{AUP_{i1}^M} \quad (8.24)$$

$$C_{ij2}^M = \frac{\Omega_{ij2}^M}{AUP_{i2}^M} \quad (8.25)$$

$$C_{ij2}^{M*} = \frac{\Omega_{ij2}^{M*}}{AUP_{i2}^{M*}} \quad (8.26)$$

In order to combine Eq. 8.24-8.26 with Eq. 8.23, the AUP^M of each type of group contribution must be corrected to the overall molecular AUP^M using the following correction factors:

$$CF_{1i} = \frac{AUP_{1i}^M}{AUP_i^M} \quad (8.27)$$

$$CF_{2i} = \frac{AUP_{2i}^M}{AUP_i^M} \quad (8.28)$$

$$CF_{2i}^* = \frac{AUP_{2i}^{M*}}{AUP_i^M} \quad (8.29)$$

Thus the cluster of the molecule is found to be the following:

$$C_{ij}^M = CF_{1i} \cdot C_{ij1}^M + CF_{2i} \cdot C_{ij2}^M + CF_{2i}^* \cdot C_{ij2}^{M*} \quad (8.30)$$

As with previous cluster algorithms, a system of three properties can be plotted on a ternary diagram. An overview of the attribute based computer aided molecular design (aCAMD) conversion algorithm is presented in Fig. 8.3. The procedure for converting attribute, property, and group data into non-dimensional property operators and property clusters for later use in a candidate generation algorithm is presented in Table 8.1.

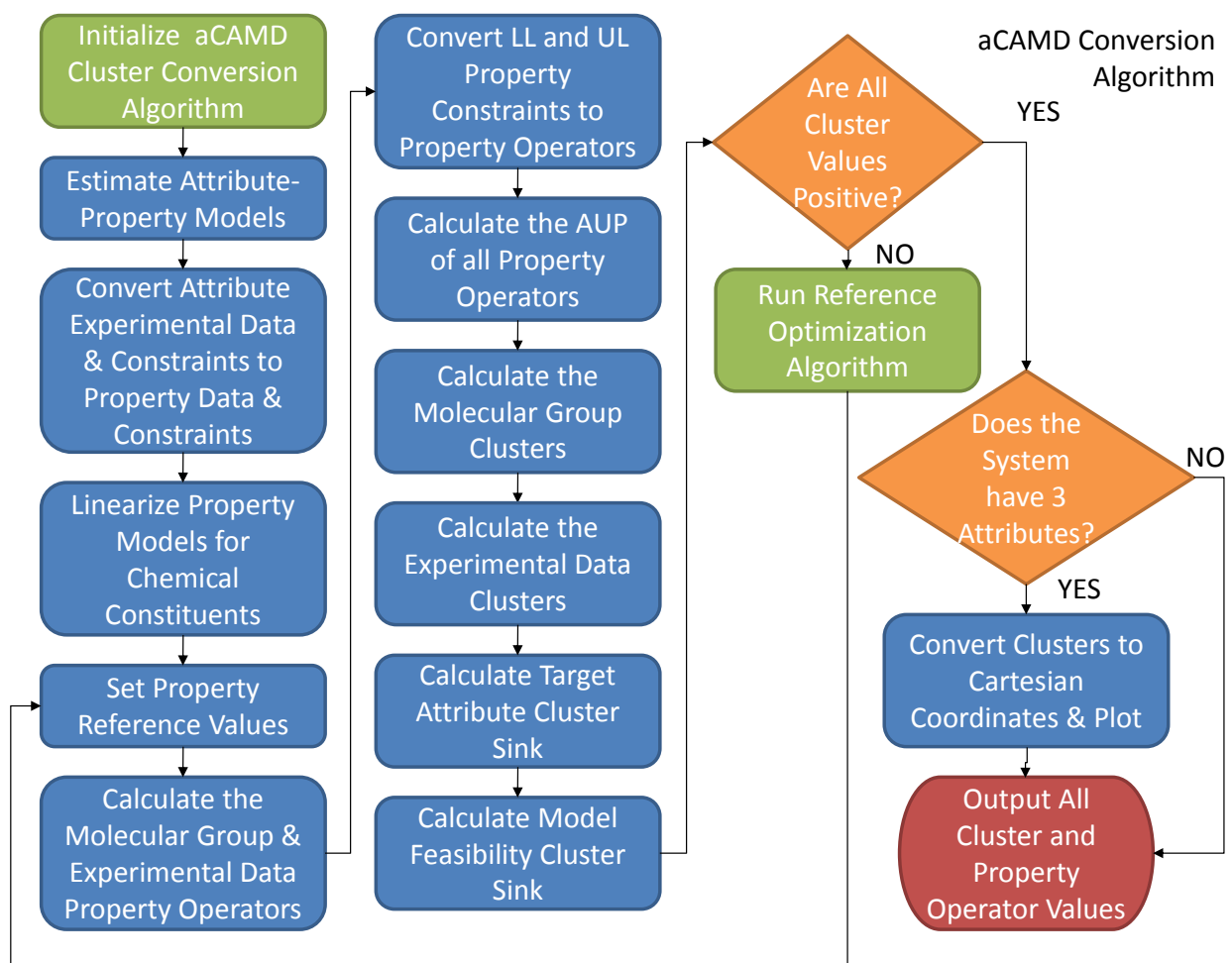


Fig. 8.3: The Attribute Based Computer Aided Molecular Design (aCAMD) Conversion Algorithm.

Table 8.1: The aCAMD Cluster Conversion Procedure

Step	Description	Equation
1	Select Attribute-Property Model	8.8
2	Fit Model to Experiment Design Point Responses to Determine Model Parameters	8.9
3	Calculate the Variance, Covariance, and Confidence Intervals of the Attribute-Property Model	8.10
4	Calculate the UL and LL of the Property Feasibility Region (pFR)	6.14, 8.10
5	Guess Property Operator Reference Values and Calculate Nondimensional Property Operators of Experimental Responses	6.4, 8.4, 8.14, 8.15

6	Calculate the AUP of the Experimental Response Property Operators	6.16
7	Calculate the Nondimensional Property Operators of the pFR	6.14, 6.15
8	Calculate the AUP range of the pFR	6.16
9	Calculate the Nondimensional 1 st and 2 nd Order Molecular Group Property Operators Found in the Chemical Constituent Training Set	6.15
10	Calculate the AUP of the Molecular Group Property Operators	6.16
11	Calculate the Maximum No. 1 st and 2 nd Order Molecular Groups that meet Step 7	6.43, 6.48
12	Calculate the 1 st and 2 nd Order Molecular Group Clusters, Experimental Response Clusters, and Cluster Domain of the pFR	6.17
13	Calculate the Variance, Covariance, and Confidence Intervals of the Linearized Property – Component Property Model	8.7
14	Calculate the UL and LL of the Property Model Feasibility Regions	6.14, 8.7
15	Calculate the Nondimensional Property Operators of the pMFR	6.14, 6.15
16	Calculate the AUP range of the pMFR	6.16
17	Calculate the Cluster Domain of the pMFR	6.17
18	Run Reference Optimization Algorithm for Non-Positive AUPs	6.8-6.12
19	Convert all Clusters to Cartesian Coordinates and Plot	6.23-6.26

A list of first and second order group contributions for the properties outlined in Table 6.1 can be found in Marrero and Gani (2001). More detail on the clustering of 2nd order and higher expressions is provided in Section 8.3. In addition, Chemmangattuvalappil *et al.* (2010) has developed third order molecular property operators for larger, more complex chemical constituents including those that are polycyclic (Marrero and Gani, 2001). After all of the molecular group property contributions have been obtained and converted to clusters, they can be used to solve a CAMD in the cluster domain. However, as was shown in Section 4.4 and Section

6.3, the structure of molecular property operators dictates that the CAMD occurs in the property operator domain. The details of the CAMD are discussed in Section 8.3.

8.3. aCAMD using Experimental Data and GCM Models

Like the aCAM^bD approach in Chapter 7, the cluster conversion algorithm for the aCAMD contains two feasibility regions. For the aCAMD these regions are the “model” feasibility region (pMFR) and the “target” feasibility region (pFR). To be considered a potential candidate, either a pure component or a mixture of components must fall within both regions and meet both *AUP* values simultaneously. Hence, this structure of the aCAMD as a reverse problem formulation (RPF) creates an opportunity to explore multiple levels of molecular architecture (e.g. group changes and pure component changes) without committing to products *a priori*, an observation that is explored more thoroughly in Chapter 9.

The procedure for finding candidate mixtures has already been discussed in Chapter 6 for known property models and in Chapter 7 for empirically derived property models. The method of finding candidate molecules is dependent upon the estimation of the molecular *AUP* and molecular clusters. For first order groups, Eq. 6.38 – 6.40 dictate that the combination of groups will obey linear mixing and lever arm analysis on a ternary cluster diagram (see Fig. 6.9). The addition of second order groups to this construct introduces non-linearity to the solution since second order groups “correct” first order group approximations by accounting for group-to-group binary interactions. The non-linearity of binary interactions prevent the use of lever-arm analysis in the clustering domain, although recent progress discussed in Chapter 10 suggests that with alteration, it may be achievable. For this situation, it has been shown by Chemmangattuvalappil *et al.* (2008) that a solution can be reached by utilizing property operators directly in an algebraic approach similar to that suggested in Section 4.4.

The general aCAMD statement is to generate the structures of all possible $i \in u$ molecules that can be built from g_l groups with p target properties. There will be one upper and one lower bound on each property estimated through the process design problem.

$$P_j^L \leq P_{ij}^M \leq P_j^U \quad j \in p, i \in u \quad (8.28)$$

Where i is the index of molecules and j is the index of properties. This equation can be written in terms of normalized property operators as follows

$$\Omega_j^L \leq \Omega_{ij}^M \leq \Omega_j^U \quad (8.29)$$

Here, Ω_{ij}^M is the p normalized j property operator of molecule i . The target property Ω_{ij}^M is described by two inequality expressions, one for the lower bound and one for the upper bound (Qin *et al.*, 2004), resulting in $2p$ inequality expressions to represent all the possible solutions (Eden *et al.* 2004; Chemmangattuvalappil *et al.* 2009; Solvason *et al.* 2009). This target property sink area is the made up of the intersection of the pFR and pMFR. Although it is possible to combine these regions into a pseudo target region by taking the minimum of the upper bounds and the maximum of the lower bounds, it is not preferred because of the role of aCAMD is both screening and optimization. Confounding the pFR and pMFR rejection criteria prevents the researcher from discerning which set of models the non-candidate solution failed and, by extension, why the non-candidate solution failed. For example, the reason a non-candidate solution fails the pFR domain check may only be because of a poorly fitted experimental model in its domain which could be solved by conducting validation experiments for a better fit. However, a violation of the pMFR domain is more severe and suggests the underlying molecular architecture is either poor or poorly fit, the latter of which is highly unlikely. In summary, the preferred approach to generating solutions that meet both the pFR and pMFR is to first check them against the pFR and then validate them using the pMFR, as was the case for CAM^bD.

In order to determine which potential candidate structures to test, the reverse problem formulation (RPF) is initiated using a similar approach suggested in Section 6.3. First the number of the set of $g_1 \in F$ first order groups found in the molecular design are calculated using Eq. 6.43 and placed in Eq. 6.48.

$$n_{g_1}^U \leq \text{Int}\left(\frac{AUP^U}{AUP_{g_1}}\right) \quad (6.43)$$

$$1 \leq n_{g_1} \leq n_{g_1}^U \quad (6.48)$$

Then the number of the set of $g_2 \in G$ second order groups are calculated from the first order groups using Eq. 8.17 and Eq. 8.19.

$$n_{ig2} = \text{Int}\left(\text{Min}\left(\frac{n_{ig1}}{\eta_{ig1}} : \frac{n_{iF}}{\eta_{iF}}\right)\right) \quad (8.17)$$

$$n_{ig2}^* = \text{Int}\left[\text{Min}\left(\frac{n_{is2}}{\eta_{g_2v_2}} : \frac{n_{iG}}{\eta_{g_2G}}\right) - \text{Min}\left(\frac{n_{ig1}}{\eta_{ig1}} : \frac{n_{iF}}{\eta_{iF}}\right)\right] \quad (8.19)$$

Once the sets are calculated, the property contributions are obtained from literature sources e.g. Marerro and Gani (2001), Constantinou and Gani (1994), etc.

Once the maximum number of dissimilar groups μ_g^{max} is set, the design is executed, beginning with a minimum combination of dissimilar groups and then increasing sequentially until the maximum is reached. For “pure group” mixtures (e.g. $\mu_g = 1$), the pure group property contributions are estimated from Eq 6.37 for each g_1 group type since pure groups have no binary interactions and, by extension, no second order groups. These clusters are then tested against the points making up the feasibility bounds of a pFR. All “pure group” property clusters (e.g. Eq. 6.17) that meet Rule 1 for the pFR are then passed on to test against the pMFR. “Pure group” clusters that meet Rule 1 for both the pFR and pMFR are then tested against Rule 2 in the same

order using Eq. 8.16. All non-candidate mixtures are discarded. The remaining “pure group” mixtures are then tested against Rule 3, 4, and 6 and non-candidate solutions are discarded. Since it has been shown by Marrero and Gani (2001) that the group contribution methods (GCM) may not predict properties accurately for open-chain polyfunctional compounds with more than four carbon atoms in the main chain, a new rule is created and the “pure component” mixtures are tested against it.

Rule 14. The number of identical functional groups in the backbone of a molecule is to be limited to a maximum of four.

If $n_{g1, BB}$ are the number of first order groups in the backbone, then Rule 14 can be written as Eq. 8.33:

$$\sum n_{g1, BB} \leq 3 \quad (8.33)$$

Pure group molecules that fail Rule 14 are discarded. The remaining groups are denoted as “pure group” candidate molecules.

For binary and larger mixtures (i.e. $\mu_g \geq 2$) of dissimilar types of groups (i.e. $v_1 \neq g_1$) with no ring compounds (i.e. $N_{RG} = 0$), the aCAMD differs from the procedure for CAMD with clusters. Since dissimilar groups are capable of forming second order groups which have negative property operators that can increase the number of groups used in the design beyond the result of Eq. 6.43, it can no longer be used as a valid bound on the design. Furthermore, an explicit set of equations to determine the maximum number of a specific first order group to take the place of Eq. 6.43 has not yet been developed. As a result, the aCAMD method for dissimilar groups defaults to a forward based approach where all combinations of dissimilar groups are generated and subsequently removed by testing against the bounds of the design. For this case

the maximum number of each group is set by the researcher, using Eq. 6.43 as a guide but not a hard bound.

Beginning with structural stability, each potential candidate solution is first tested for structural stability using Rule 4 and Rules 14-16. Solutions that fail any of these rules are discarded. Of the remaining structurally sound molecules, those solutions that fail the non-dimensional property operator bounds of the pFR are discarded, followed by those that fail the pMFR. Alternatively these could be combined into a single region described by Eq. 8.29 and tested as discussed earlier. Those solutions that remain are outputted as candidate solutions to the design problem.

The same methodology can be followed to identify cyclic (i.e $N_{RG} = 1$) compounds as well with a few additional constraints. The group contribution method has questionable range of accuracy for compounds containing more than one ring even with the inclusion of second order groups (Marrero and Gani 2001; Chemmangattuvalappil *et al.* 2009; Solvason *et al.* 2009). So the number of rings in the compound must be restricted to one. The decision on the groups to be part of the ring must be made ahead of design. This is to account for the difference in the property contributions of the same group to cyclic and acyclic compounds. No first order acyclic groups should be found in the ring structures.

Rule 15. Ring structures within a potential candidate molecule should consist of only first order cyclic groups.

Furthermore, the number of cyclic groups in a ring structure is also limited. Two-group structures are infeasible and three-group structures are often thermodynamically unstable and should be avoided. Therefore a new Rule 16 is created to check that stability of suitable candidate structures.

Rule 16. The minimum number of functional groups in a ring structure of a molecule is three.

If $n_{g1, RG}$ are the number of groups in a ring, then Rule 15 can be represented using Eq. 8.34:

$$\sum n_{g1, RG} \geq 3 \quad (8.34)$$

Ring compounds that fail Rule 15 and 16 are discarded. Those that meet Rule 15 and 16 as well as Rule 1-4 are denoted candidate solutions.

In conclusion, this approach represents a significant reduction in the complexity of an aCAMD. For example, the theoretical 2050+ parameters of the earlier aCAM^bD design problem can be reduced to only 20 parameters in an aCAMD, a 100:1 reduction. This reduction in complexity may also improve the accuracy of the design by: (1) using the additional degrees of freedom in the design to minimize the error and (2) ensuring each enumerated candidate molecular structure possesses the desired values for the consumer attribute. The only caveat is that accuracy of the aCAMD is predicated on the quality of two models (i.e. the attribute-property relationship and the property-molecular architecture relationship), whereas the CAM^bD utilizes a single model. Combining the predicted fit Q_p^2 of the two models using Eq. 3.26, and combining them using Eq. 8.35:

$$Q_{z^M}^2 = \prod_{z=1}^{z^M} (Q_z^2)^{\frac{1}{z^M}} \quad (8.35)$$

where z^M is the total number of models used in the design, and z is the model identifier. Values of Eq. 8.35 that are close to 1.0 would indicate a completely validated molecular design containing only global optimum candidates, whereas values less than 1.0 suggests that some error may exist and that the data set may contain non-global optimum candidate solutions. In general, a larger value of the $Q_{z^M}^2$ measure indicates a better chance that the estimated global

optimum candidate set is the true global optimum candidate set. In situations where Q_{zM}^2 is of questionable accuracy, a second validation check can be executed by directly calculating the attributes for the candidate molecules and checking them against the bounds of the design given in Eq. 8.11. The calculation of the candidate attributes is achieved by solving the attribute-property model with a NLP solver of the inverse regression problem or a direction calculation from the normal regression problem. Each additive attribute can then be plotted to verify that the solution meets the specified consumer attributes.

Since the structure of the proposed method does not automatically require the solution of an NLP, it can be certain that the potential additives will always be at the estimate of the global optimum. The same cannot be said for the solution of most MINLP algorithms. Although the MINLP will always provide a solution, it may only be a local optimum. There is no way of knowing if the local optimum is close to the global optimum without using more rigorous and computationally expensive solution strategies like disjunctive programming. Using the method outlined in this chapter provides not only estimates of the global optimum, but also provides a means by which to determine how well the estimated global optimum matches the true global optimum, all while being computationally inexpensive.

The procedure determining finding the property targets from a set of consumer attributes and solving for candidate mixtures and molecules that meet the design targets is summarized in Table 8.3. and Fig. 8.4 – Fig. 8.6.

Table 8.3: The Attribute Based Computer Aided Molecular Design (aCAMD) Method.

Step	Description	Equation
1	Set Maximum Dissimilar Groups in the Mixture	-

2	Calculate Pure Group Solution Clusters	6.37, 6.16, 6.40
3	Discard Pure Group Clusters (from aCAMD Conversion Algorithm) that Fail Rule 1 for the pFR	8.28, 8.10, 8.15
4	Discard Pure Group Clusters (from aCAMD Conversion Algorithm) that Fail Rule 6 for the pMFR	8.28, 8.7, 8.15
5	Calculate No. Pure Groups	6.47, 6.48
6	Calculate Pure Group Solution AUPs, Discard Pure Group Solutions that Fail Rule 2 and Rule 3 for the pFR	8.28, 8.10, 6.16, 6.47
7	Discard Pure Group Solutions that Fail Rule 7 and Rule 8 for the pMFR	8.28, 8.7, 6.16, 6.42
8	Discard Pure Group Solutions that Fail Rule 4 and Output Candidate Solutions	6.41
9	Repeat Steps 2-7 for Ring Structures and Output Candidate Solutions	-
10	For Binary+ Solutions, Generate All Group Combinations Given a Maximum Number of Each 1 st Order Group	6.46, 6.48
11	Discard Binary+ Solutions that Fail Rule 4	6.41
12	Discard Binary+ Solutions that Fail Rule 14	8.33
13	Calculate the No. 2 nd Order Groups	8.17, 8.19
14	Calculate the Property Operator Responses for the Candidate Solutions	8.23, 8.22, 8.20, 8.15
15	Discard Binary+ Solutions that Fall Outside the pFR	8.28, 8.10
16	Discard Binary+ that Fall Outside the pMFR, Output Candidate Solutions	8.28, 8.7
17	Repeat Steps 10-14 for Ring Structures	-
18	Discard Binary+ Solutions that Fail Rule 15	-
19	Discard Binary+ Solutions that Fail Rule 16	8.34
20	Discard Binary+ Solutions that Fall Outside the pFR	8.28, 8.10
21	Discard Binary+ Solutions that Fall Outside the pMFR, Output Candidate Solutions	8.28, 8.7
22	Repeat Steps 10-21 until μ^{max} is reached	-

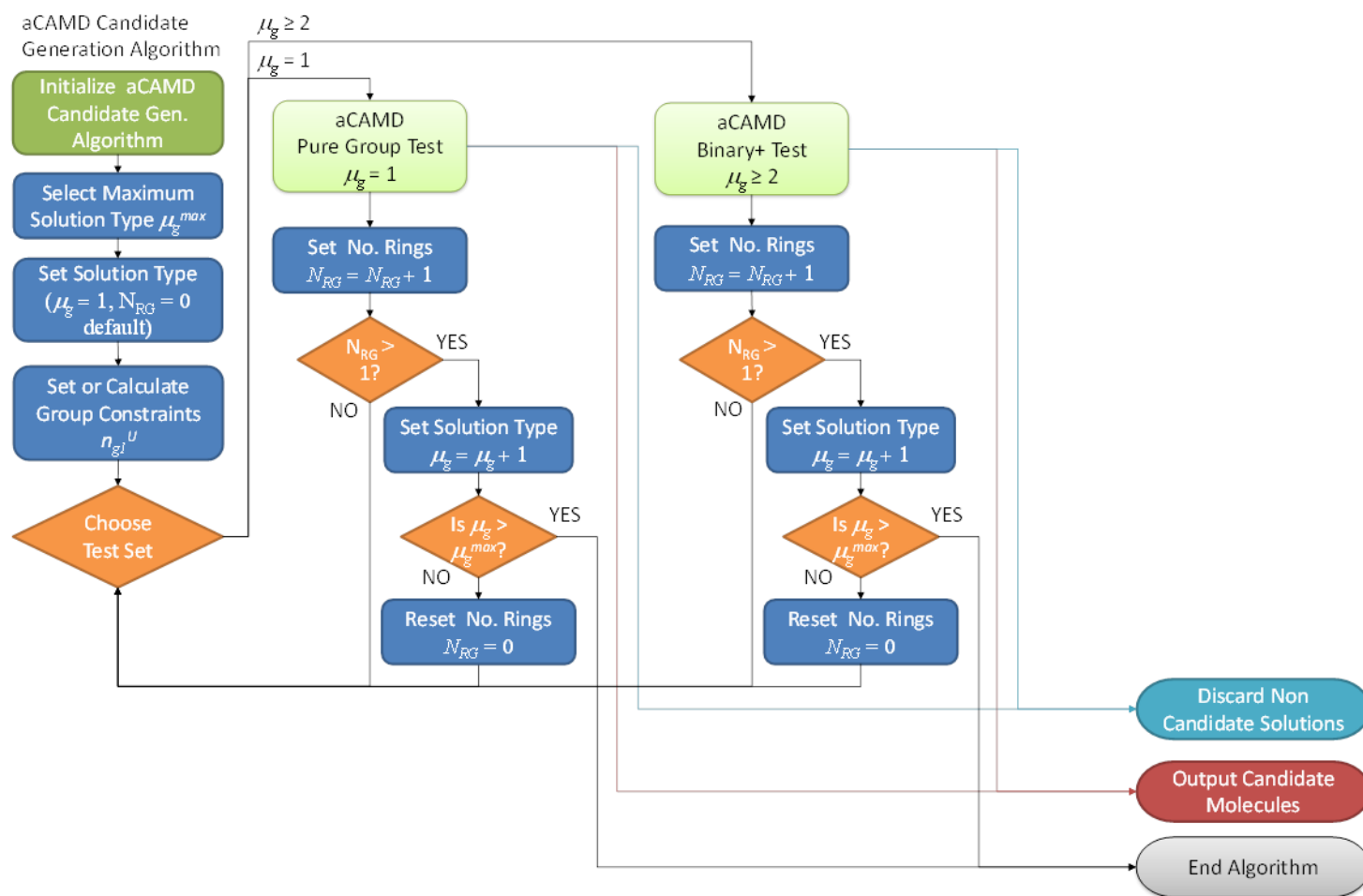


Fig. 8.4: The aCAMD Candidate Generation Algorithm

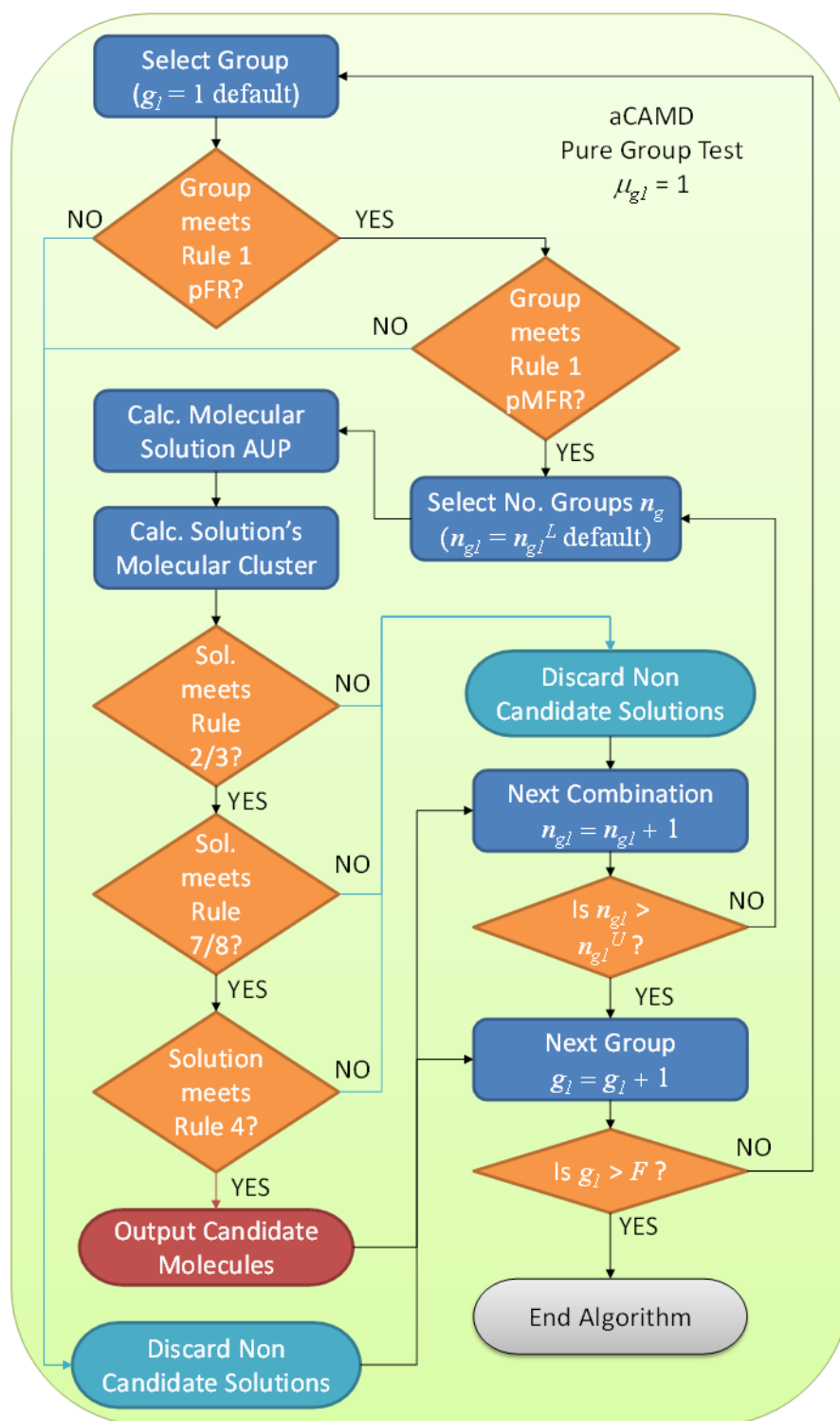


Fig. 8.5: The aCAMD Pure Group Test Candidate Molecules



Fig. 8.6: The aCAMD Binary + Test for Candidate Molecules

8.4. Case Study: Environmentally Friendly Refrigerant Design

Since the 1930s chlorofluorocarbon (CFC) based refrigerants, such as dichlorofluoromethane (R-12), have found use in home refrigerators, aerosol cans, fire extinguishers, and automotive air conditioners primarily due to their non-toxic, non-flammable nature and their high overall thermodynamic efficiency (He *et al.*, 2005). Unfortunately, R-12 has been found to damage the ozone layer, increase greenhouse effects, damage telluric environments, and affect human health (Zhao *et al.*, 2004). Beginning in 1987 with the Montreal Protocol on substances that deplete the ozone layer, CFC and hydrochlorofluorocarbon (HCFC) refrigerants were phased out and replaced with more ozone friendly hydrofluorocarbon (HFC) refrigerants such as 1,1,1,2-tetrafluoroethane (R-134a). Although R-134a has been easily adapted to fit existing refrigeration cycles, it exhibits a high global warming potential (*GWP*) leading to a moderately high potential environmental impact (*PEI*). In 1997 the Kyoto Protocol on reducing greenhouse gases identified R-134a as contributing to global warming (Zhao *et al.*, 2004). Thus, the use of R-134a is only a temporary solution to the search for environmentally friendly refrigerants capable of meeting the thermodynamic needs of refrigerant systems.

The design of these environmentally friendly replacement refrigerants has primarily proceeded down one of two paths: either mixing existing refrigerants experimentally (Akasaka *et al.*, 2007; Coquelet *et al.*, 2006; Eljack *et al.*, 2005; Kondo *et al.*, 2006; Zhao *et al.*, 2004) or computationally (Saleh and Wendland, 2006; Calero *et al.*, 1998; Duvedi and Achenie, 1997; Molnarne *et al.*, 2005; Muller *et al.*, 1996; Rachidi *et al.*, 1997; Sahindis *et al.*, 2003; Schroder and Molnarne, 2005; Young and Cabezas, 1999). Using the experimental method, He *et al.* (2005) have determined that using a 85/15 wt% of difluoroethane (R-152a) and pentafluoroethane (R-125) is a potential substitute for dichlorofluoroethane (R-12) in a typical

home refrigerator. Others such as Kondo *et al.* (2006) have focused on improving the flammability issues associated with using propane (R-290). Although these traditional experimental techniques offer valuable information to the design of refrigerant mixtures, it is time-consuming and expensive to test all possible candidates. Use of a mixed integer non-linear program (MINLP) to elucidate a computer-aided molecular design (CAMD) offers a unique approach to the candidate enumeration problem. As shown by Duvedi and Achenie (1997), the MINLP is limited to local optima solutions when the property prediction methods are non-linear. In both the experimental and MINLP techniques there is no guarantee the global optimum exists among the list of candidate molecules. Recently, Sahindis *et al.* (2003) utilized a branch-and-bound global optimization algorithm in a typical MINLP as a method to enumerate all feasible candidate molecules. While a significant contribution to the improvement of the MINLP, the method still relies on several intensive algorithms, which are computationally expensive. Furthermore, it offers no additional insights to solution of the problem such as the relative impacts of each of the candidate molecules and molecular groups, their similarities, and their interactions.

In addition, in product design, consumer preference drives the value of a product. If the consumer does not prefer the product to similarly priced products, then the product is not selected. To measure the consumer preference, a set of consumer attributes $\mathbf{A}$ are defined. Often, the relationship between the underlying properties and the consumer attribute is not explicitly known. In these situations, design of experiments is employed to find approximate empirical relationships between the attributes and either the chemical constituents or the underlying properties. For refrigeration systems the attributes important to the consumer are the Energy Index (EI), the Compressor Displacement Index (CDI), the Miscibility Index (MI), and

the Potential Environmental Impact (*PEI*). The *EI* is a measure of the thermodynamic efficiency of the refrigerant in an ideal refrigeration loop with a condenser temperature of 300K and an evaporator temperature of 233K. The *CDI* is measure of the compressor displacement. The *MI* is a measure of the compatibility of the designed refrigerant with the existing equipment. The *PEI* is the potential environmental impact of the designed refrigerant or refrigerant mixture as calculated by the WAR algorithm (Young and Cabezas 2000). The *PEI* is an estimated value consisting of weighted measures of the toxicology, ozone depletion potential (*ODP*), global warming impact (*GWP*), and many other factors important to health and the environment. All *PEI* values estimated were done using the default weighting factors provided by WAR GUI Version 1.0.14 (Young and Cabezas 2000).

As noted earlier, R-134a is an excellent thermodynamic refrigerant, but its *PEI* value is too high because of a high *GWP*. In this case study it is desired to find a suitable refrigerant with a low *PEI* number to replace R-134a. In addition, the customer already has access to a low-cost source of difluoromethane (R-32) and wishes to find an additive that can be added to it so that the combined mixture can replace R-134a. Based on commercially available refrigerants and proprietary knowledge, the customer has recommended using one of the 7 refrigerants listed in Table 8.4 as the additive, although the use of other additives remains possible.

Table 8.4: List of Potential Refrigerant Additives

ASHRAE #	Name	CAS #
R-152a	1,1 – Difluoroethane	75-37-6
R-290	Propane	74-98-6
R-600	Butane	106-97-8

R-125	Pentafluoroethane	354-33-6
R-143	1,1,2-Trifluoroethane	430-66-0
R-22	Chlorodifluoromethane	75-45-6
R-600a	Isobutane	106-97-8

A set of acceptable *EI*, *CDI*, *MI*, and *PEI* values have also been developed that are deemed by the customer as suitable for a drop-in replacement of R-134a. They are as follows:

$$40 \leq EI \leq 55 \quad (8.36)$$

$$24 \leq CDI \leq 28 \quad (8.37)$$

$$1.8 \leq MI \leq 2.2 \quad (8.38)$$

Since the *PEI* cannot be estimated by the GCM, it will be used as an additional screening tool for the pure component and candidate mixtures found to meet the other attribute targets such that the candidate solution with the smallest *PEI* is chosen as the output of the aCAMD. Thus, the overall objective of this case study is to generate the structures of all possible additive and additive mixtures that can be built from N_{g1} groups found in the customer's training set that, when combined with R-32, can be utilized as a replacement for R-134a.

Two methods exist in which these candidate pure components and mixtures can be ascertained. The first method is to conduct a mixture design of experiment (MDOE) and fit the attribute response data in terms of chemical contributions directly using the procedure outlined in Chapter 7. For this method, the additives listed in Table 8.4 are combined with the refrigerants R-12, R-134a, and R-32 and a d-optimal MDOE is conducted. Using the recommendations from Custom Designer JMP 7.0.2, 31 design points of various mixture fractions of the 10 components were developed as shown in Table 8.5.

Table 8.5: The MDOE for the Refrigerant Replacement Case Study.

Run	Mass Fractions									
	R-32	R-152a	R-290	R-600	R-125	R-134a	R-143	R-12	R-22	R-600a
1	0.00	0.00	0.00	0.00	0.01	0.99	0.00	0.00	0.00	0.00
2	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.00
3	0.00	0.05	0.04	0.00	0.91	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.02	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
6	0.00	0.00	0.00	0.00	0.89	0.00	0.11	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	0.10	0.00
8	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.01	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
12	0.00	0.00	0.90	0.00	0.00	0.09	0.00	0.01	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
14	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
16	0.00	0.07	0.61	0.00	0.00	0.00	0.32	0.00	0.00	0.00
17	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
20	0.77	0.00	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.00
21	0.00	0.00	0.00	0.00	0.00	0.85	0.12	0.02	0.01	0.00
22	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
23	0.00	0.00	0.00	0.10	0.00	0.90	0.00	0.00	0.00	0.00
24	0.36	0.00	0.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25	0.00	0.00	0.00	0.02	0.00	0.00	0.98	0.00	0.00	0.00
26	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00
27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
28	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
29	0.00	0.87	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00
30	0.10	0.00	0.00	0.00	0.90	0.00	0.00	0.00	0.00	0.00
31	0.07	0.00	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00

The associated attribute responses of the 31 mixture fractions provided by the theoretical customer are given in Table 8.6.

Table 8.6: The Attribute Responses to the Refrigerant Replacement MDOE.

Run	Attribute Responses			
	<i>EI</i> (Measured)	<i>CDI</i> (Measured)	<i>MI</i> (Measured)	<i>PEI</i> (Calculated)
1	51.7	26.4	1.99	0.79
2	58.9	31.4	1.76	1.34
3	56.5	23.9	1.69	1.05
4	100.0	26.3	3.65	0.87
5	55.9	30.9	1.73	1.49
6	54.4	23.6	1.58	1.13
7	37.5	26.3	1.57	2.63
8	71.1	28.8	2.82	0.05
9	66.0	24.0	4.27	1.28
10	67.6	24.6	2.05	1.24
11	69.1	23.0	4.10	1.47
12	95.9	26.2	3.54	0.86
13	69.1	23.0	4.10	1.47
14	90.1	36.0	1.96	0.00
15	35.7	25.7	1.55	2.75
16	89.5	25.9	3.13	0.92

17	71.1	28.7	2.82	0.05
18	69.1	23.0	4.10	1.47
19	55.9	30.9	1.73	1.49
20	80.0	33.9	1.98	0.18
21	52.8	26.2	1.99	0.89
22	66.2	24.0	4.29	1.27
23	53.7	26.2	2.20	0.83
24	97.3	29.8	2.93	0.54
25	67.5	24.6	2.09	1.24
26	83.1	35.0	1.94	0.28
27	35.7	25.7	1.55	2.75
28	66.2	24.0	4.30	1.27
29	68.9	28.5	2.72	0.14
30	57.0	24.8	1.59	1.00
31	69.2	25.4	2.04	1.16

These responses, targets, and fitted models were then non-dimensionalized using reference property operators of 10 for *EI*, 3 for *CDI*, and 0.2 for *MI* and clustered using the procedure outlined in Chapter 7, resulting in Fig. 8.7.

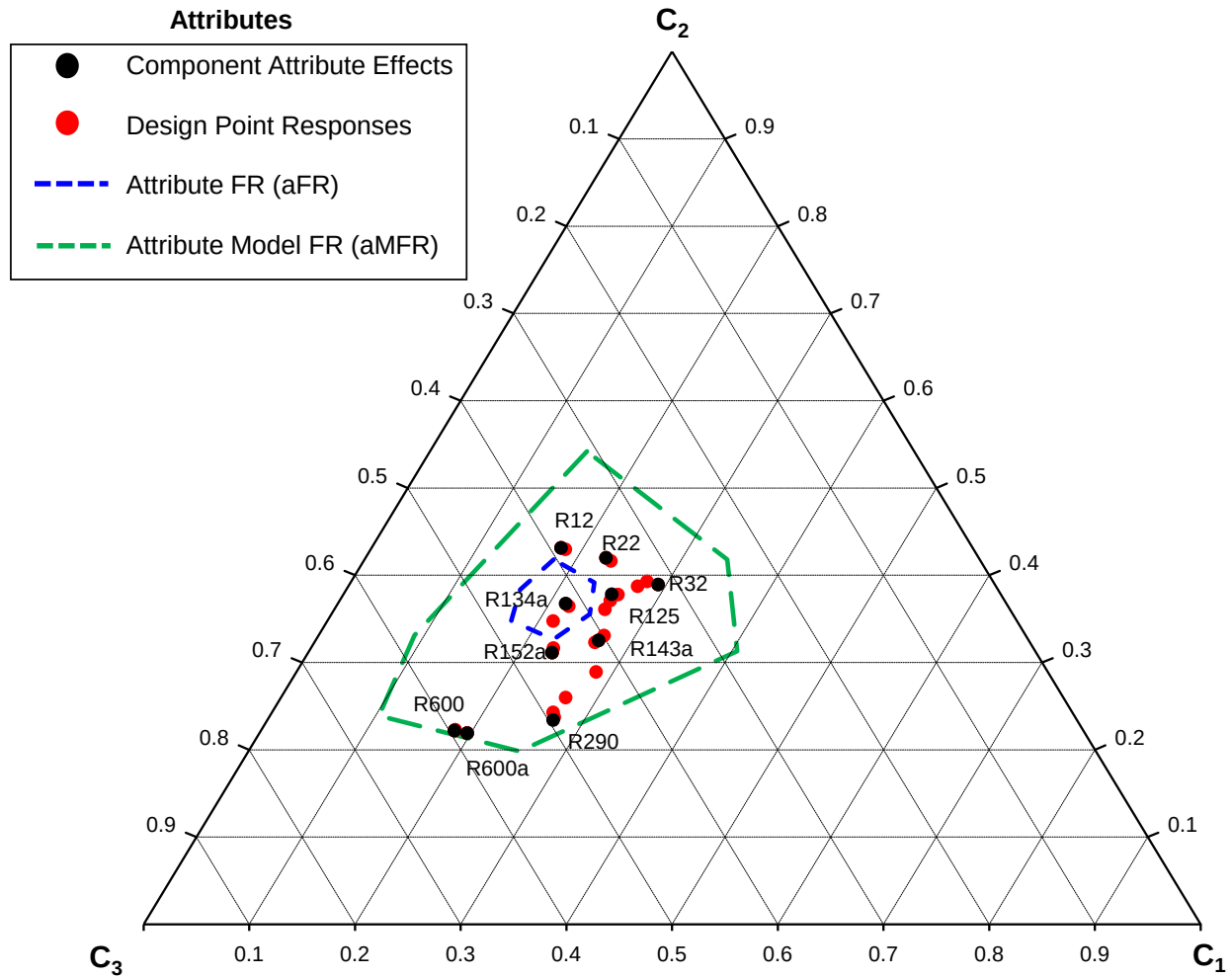


Figure 8.7: The Attribute Cluster Diagram for Refrigerant Replacement Case Study.

As shown, only R-134a meets the target attributes. No other single component meets the necessary target attributes, a result common in refrigeration systems (Zhao *et al.* 2008). In addition, although some of the designed mixtures fall within the target feasibility region, no tested mixture satisfied Rules 1-3 simultaneously. It is also clear that no new mixtures of the candidates suggested by the customer with R-32 will satisfy the target region when using lever-arm analysis, unless that candidate is also mixed with R-134a. This option was deemed unacceptable since the customer desired to fully replace R-134a and, as a result no mixture of the components given by the customer will produce a solution to this design problem.

Rather than attempt to augment the mixture design with additional chemical constituents, new molecular structures were designed from the underlying molecular architecture present in the design using the aCAMD approach discussed earlier in this chapter. In this approach the attribute data is mapped to a property domain where group contribution models can be used to build new molecules not present in the original training set using the training set's functional groups. The attribute relationship described by Eq. 8.5 must be defined such that the number of attributes is equal or greater than number properties used in the design (e.g. $a = 3$ means $p \geq 3$). There is reason to believe that EI , CDI , and MI are functions of the following properties: critical temperature T_c , critical pressure P_c , and the latent heat of vaporization at standard conditions H_v . All of these properties can be described by GCM. Furthermore it has been found that EI is proportional to T_c , P_c , and H_v , CDI to T_c and P_c , and MI to T_c , P_c , and H_v . The simplest model that can be built from these intuitive relationships is that of first order polynomials as shown in Eq. 8.39 – 8.41:

$$EI = \beta_o + \beta_1 \cdot T_c + \beta_2 \cdot P_c + \beta_3 \cdot H_v \quad (8.39)$$

$$CDI = \beta_o + \beta_1 \cdot T_c + \beta_2 \cdot P_c \quad (8.40)$$

$$MI = \beta_o + \beta_1 \cdot T_c + \beta_2 \cdot P_c + \beta_3 \cdot H_v \quad (8.41)$$

To determine the regressors in these relationships, property data for each of the design point mixtures was provided by the customer as shown in Table 8.7. These properties could also be determined using equations-of-state and group contribution property estimations.

Table 8.7: The Calculated Property Responses to the Refrigerant Replacement MDOE (Marrero and Gani 2001).

Run	Properties		
	T_c (K)	P_c (bar)	H_v (kJ/mol)
1	373.8	40.5	18.1
2	367.5	50.7	15.6
3	342.9	36.7	13.5
4	369.8	42.7	14.6
5	369.3	49.9	15.8
6	340.2	36.2	13.2
7	383.4	42.2	16.6
8	386.4	45.2	18.4
9	425.1	38.0	20.8
10	346.8	38.3	13.5
11	408.2	36.5	19.1
12	370.3	42.3	14.9
13	408.2	36.5	19.1
14	351.4	57.9	14.2
15	385.0	41.3	16.7
16	363.7	41.4	14.5
17	386.4	45.2	18.4
18	408.2	36.5	19.1
19	369.3	49.9	15.8
20	356.6	53.9	15.1
21	371.2	40.4	17.6

22	425.3	38.0	20.8
23	379.4	40.3	18.5
24	363.1	48.1	14.4
25	348.3	38.3	13.6
26	354.8	56.3	14.4
27	385.0	41.3	16.7
28	425.4	38.0	20.8
29	384.9	44.6	18.4
30	340.6	38.2	13.2
31	347.1	39.6	13.5

The regressors found to fit the above equations were estimated using the statistical package JMP 7.0.2 and are listed in Table 8.8.

Table 8.8: The Regressors for the Refrigerant Attribute Models

Informative Statistics	All Data Points ($n = 31$)			Excluding high R12 ($n = 28$)		
	<i>EI</i>	<i>CDI</i>	<i>MI</i>	<i>EI</i>	<i>CDI</i>	<i>MI</i>
β_0	-116.3887	5.1674	-12.7938	-209.2422	5.0184	-16.8315
β_1	0.6467	-0.0070	0.0551	1.0933	-0.0064	0.0745
β_2	1.4505	0.5726	-0.0085	1.4181	0.5720	-0.0100
β_3	-7.3196	-	-0.2985	-11.5491	-	-0.4825
R^2	0.3015	0.9902	0.5989	0.5611	0.9915	0.8723
Q^2	0.1615	0.9891	0.5246	0.4434	0.9908	0.8379
R^2_{adj}	0.2239	0.9895	0.5543	0.5062	0.9908	0.8564

When performing the regression, it is apparent that a poor fit for *EI* is achieved with an R^2 value of 0.3015, due primarily to the experiments containing large fractions of

dichlorofluoromethane (R-12). It may be beneficial to remove these experiments from the design should they be determined to be outliers. Continuing with the other attribute data, the *CDI* is found to fit the data well with a R^2 value of 0.9902. The *MI* is also found to fit the data poorly, with an R^2 value of 0.5989. Again, experiments containing large fraction of R-12 appear to be outliers. An inspection of actual vs. predictive plots also suggests that these experiments might be outliers. In Figure 8.7a the lines of best fit and the confidence intervals are plotted using all experiments. Figure 8.7b shows the same information as Fig. 8.7a with the suspected outliers removed. To determine if potential outliers should be excluded, an analysis is performed to determine if the experiments with large amounts of R-12 overly influenced the model. In this method, the suspected experiments are removed and the regression is repeated. The results are compared to the prior results containing all possible solutions. If large changes in the R^2 and the regressors are observed, then the experiments are considered overly influential and removed from the data set. In Fig. 8.7b it is found that the R^2 values were significantly improved. The R^2 values for *EI*, *CDI*, and *MI* improved to 0.5611, 0.9915, and 0.8723, respectively. The parameter estimates also changed noticeably. Thus it is concluded that the experiments involving large amount of R-12 were overly influential in determining the regressors and are removed from the analysis. When removing the experiments containing large amounts of R-12 from design, the range of the model must also be adjusted by placing boundaries around the region of measured values only. This result is an acceptable limitation since the R-12 *PEI* is the highest of the compounds evaluated.

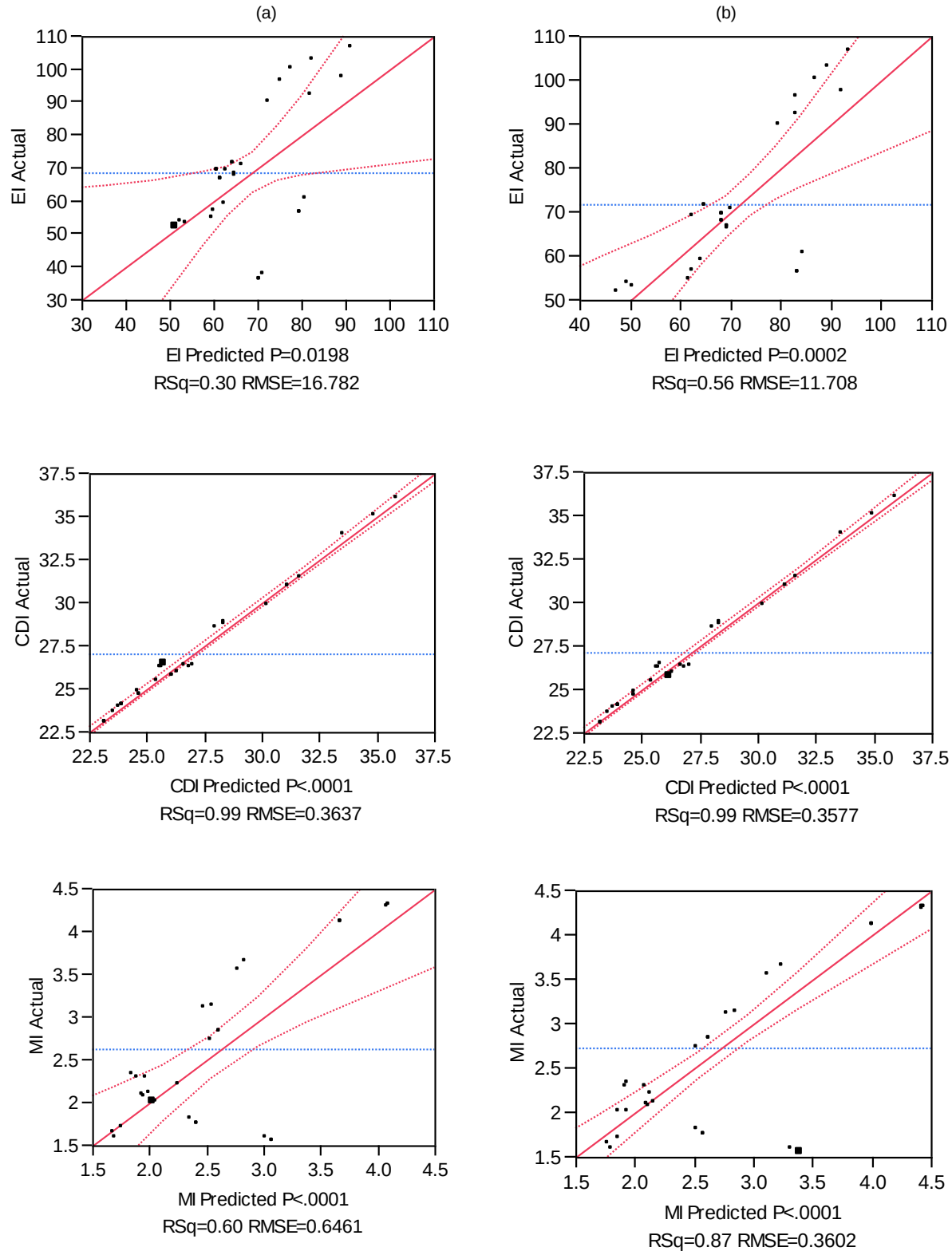


Figure 8.8: The actual vs. predicted plots of EI , CDI , and MI with (a) all of the design points ($n=31$) and (b) the high level R-12 experiments removed ($n=28$).

Further improvement may also be obtained by developing a non-linear relationship for Eq. 8.5. In this case, a non-linear solver must be used to find the optimum set of properties that deliver the set of attributes. Unfortunately, the use of non-linear programming introduces increased computational complexity and the potential for returning non-global optimum solutions. An advantageous, alternate technique for building stronger attribute-property relationships is to use inverse regression. Inverse regression reverses the dependent and independent variables in Eq. 8.5 to arrive at Eq. 8.42:

$$P = f(A) \quad (8.42)$$

The inverse linear regression problem is only valid when both the properties and attributes are random variables meaning they are normally distributed. This is true for this case study since the attributes are system responses and the properties follow a normal curve. The resulting inverse regression expressions for T_c , P_c , and H_v are shown in Eq. 8.43-8.45:

$$T_c = \beta_o + \beta_1 \cdot EI + \beta_2 \cdot CDI + \beta_3 \cdot MI \quad (8.43)$$

$$P_c = \beta_o + \beta_1 \cdot EI + \beta_2 \cdot CDI + \beta_3 \cdot MI \quad (8.44)$$

$$H_v = \beta_o + \beta_1 \cdot EI + \beta_3 \cdot MI \quad (8.45)$$

The regressors for these expressions are shown in Table 8.9.

Table 8.9: The Regressors for the Inverse Regression Expressions.

Informative Statistics	Excluding high R12 ($n = 28$)		
	T_c	P_c	H_v
β_0	266.3071	-5.7126	16.1679
β_1	-1.1867	0.0105	-0.0834
β_2	3.5923	1.7235	-
β_3	35.0855	0.3916	2.2703
R^2	0.9917	0.9946	0.7243
Q^2	0.9892	0.9925	0.6624
R^2_{adj}	0.9906	0.9939	0.7022

Using the expressions in Eq. 8.43-8.45, the upper and lower bounds for each consumer attribute listed in Eq. 8.36-8.38 were converted into a set of minimum and maximum limits on the physical-chemical properties, as shown in Table 8.10.

Table 8.10: Upper and Lower Limits on the Physical-Chemical Properties Estimated.

FR	T_c	P_c	H_v
LL	350.4	36.9	15.7
UL	396.6	43.8	17.8

Since some uncertainty exists in the attribute-property models, new lower and upper bounds are estimated from these minimum and maximum limits using Eq. 8.10. This broadening of the feasibility range ensures that, with 95% confidence, any molecule found outside the region is a non-candidate solution and should be discarded. The resulting target domain is known as the property feasibility region (pFR) and shown in Table 8.10.

Table 8.11: The Upper and Lower Limits on the Physical-Chemical Properties in the pFR

pFR	T_c	P_c	H_v
<i>LL</i>	342.2	35.2	11.1*
<i>UL</i>	405.7	45.6	22.8

*11.733 is the lower limit of the Marrero and Gani GCM method

Of note is the large broadening of the standard enthalpy of vaporization H_v resulting in an infeasible lower limit. This value was replaced by the lower limit of the GCM model at 11.733 kJ/mol. Likewise a constraint on the applicability range of the property-component model termed model feasibility region is mapped to the design space as the upper and lower bounds on the property domain. The region is determined using the minimum and maximum of the responses at the design points resulting in Table 8.12.

Table 8.12: The Upper and Lower Limits on the Modeled Physical-Chemical Property Responses.

MFR	T_c	P_c	H_v
<i>LL</i>	340.2	36.2	13.2
<i>UL</i>	425.4	57.9	20.8

A pMFR can be constructed by accounting for the uncertainty in the physical-chemical property response model. This uncertainty can either be estimated using Eq. 8.7 or by using the information on root mean square error (RMSE) from Marrero and Gani (2001). The RMSE for the experimental design is smaller so the first option was selected, resulting in the pMFR given in Table 8.13.

Table 8.13: The Upper and Lower Limits on the Physical-Chemical Properties in the pMFR

pMFR	T_c	P_c	H_v
LL	334.4	30.4	4.4*
UL	431.3	66.3	26.7

*11.733 is the lower limit of the Marrero and Gani GCM method

For a candidate or candidate mixture to be considered valid, it must fall within both the feasibility region and the model feasibility region and satisfy the *AUP* ranges of both regions simultaneously. The estimation of new mixtures, those not present in the original training set, occurs through the use of the aCAM^bD method which linearly mixes various pure components to arrive at the targets in Table 8.10-8.13. The pure component effects are estimated from literature sources and provided by the customer. They were corroborated using Perry and Green (1997) and Sahinidis *et al.* (2005) resulting in Table 8.14.

Table 8.14: Pure Component Properties of the Pure Component Effects

Pure Component i	ASHRAE #	Pure Component Properties		
		T_c (K)	P_c (bar)	H_v (kJ/mol)
1	R-32	351.4	57.9	11.99
2	R-152a	386.4	45.2	18.41
3	R-290	369.8	42.5	14.60
4	R-600	425.4	38.0	20.81
5	R-125	339.4	36.0	13.14
6	R-134a	374.2	40.6	18.20
7	R-143a	346.8	38.3	13.49
8	R-12	385.0	41.3	16.67

9	R-22	369.3	49.9	15.75
10	R-600a	408.2	36.5	19.08

All data necessary for the execution of the aCAMD has now been gathered, but requires conversion to nondimensional property operators, clusters, and AUP values in order to facilitate a solution. Beginning with the training set, the linearized forms of critical temperature, critical pressure, and the standard enthalpy of vaporization listed in Table 6.1 are used to convert the training set into property operators. The property operators are then nondimensionalized according to Eq. 8.15 using the reference values listed in Table 8.15.

Table 8.15: Nondimensionalized Property Operators of the Training Set

Run	Nondimensional Property Operators		
	Ω_1	Ω_2	Ω_3
1	1.01	1.22	0.64
2	0.98	0.81	0.36
3	0.88	1.43	0.17
4	0.99	1.12	0.29
5	0.99	0.84	0.40
6	0.87	1.46	0.14
7*	1.05	1.14	0.48
8	1.06	1.01	0.67
9	1.26	1.35	0.90
10	0.90	1.34	0.18
11	1.17	1.44	0.73
12	0.99	1.14	0.32

13	1.17	1.44	0.73
14	0.91	0.60	0.03
15*	1.06	1.18	0.49
16	0.96	1.18	0.28
17	1.06	1.01	0.67
18	1.17	1.44	0.73
19	0.99	0.84	0.40
20	0.94	0.71	0.17
21	1.00	1.23	0.59
22	1.26	1.35	0.91
23	1.03	1.23	0.67
24	0.96	0.90	0.19
25	0.90	1.34	0.19
26	0.93	0.64	0.07
27*	1.06	1.18	0.49
28	1.26	1.36	0.91
29	1.06	1.04	0.66
30	0.87	1.35	0.13
31	0.90	1.27	0.17
ref	5	0.05	10

*Experiments not included in the parameterization of Eq. 8.42

The resulting property operators are then be converted into property clusters using Eq. 6.6. and shown in Table 8.16.

Table 8.16: Clusters and *AUPs* of the Training Set

<i>Run</i>	Property Clusters			<i>AUP</i>
	C_1	C_2	C_3	
1	0.351	0.426	0.223	2.87
2	0.455	0.376	0.169	2.16
3	0.355	0.576	0.069	2.48
4	0.413	0.467	0.121	2.40
5	0.443	0.376	0.180	2.23
6	0.352	0.589	0.059	2.47
7*	0.392	0.427	0.181	2.68
8	0.388	0.370	0.243	2.74
9	0.358	0.385	0.257	3.52
10	0.372	0.555	0.073	2.41
11	0.349	0.431	0.220	3.35
12	0.405	0.464	0.131	2.45
13	0.349	0.431	0.220	3.35
14	0.596	0.388	0.017	1.53
15*	0.386	0.433	0.180	2.74
16	0.398	0.487	0.115	2.42
17	0.387	0.370	0.243	2.75
18	0.349	0.431	0.220	3.35
19	0.443	0.376	0.180	2.23
20	0.516	0.391	0.093	1.81
21	0.354	0.437	0.208	2.81
22	0.357	0.385	0.258	3.52

23	0.351	0.420	0.229	2.94
24	0.468	0.438	0.093	2.05
25	0.371	0.550	0.078	2.43
26	0.565	0.391	0.044	1.64
27*	0.386	0.433	0.180	2.74
28	0.357	0.385	0.258	3.52
29	0.383	0.376	0.241	2.76
30	0.372	0.573	0.055	2.35
31	0.385	0.544	0.071	2.33

Likewise, the nondimensionalized property operators and clusters for each pure component effect are also calculated and shown in Table 8.17 and Table 8.18, respectively.

Table 8.17: Nondimensionalized Property Operators of the Pure Component Effects

ASHRAE #	Nondimensional Property Operators		
	Ω_1	Ω_2	Ω_3
R-32	0.91	0.60	0.03
R-152a	1.06	1.01	0.67
R-290	0.99	1.13	0.29
R-600	1.26	1.36	0.91
R-125	0.87	1.47	0.14
R-134a	1.01	1.22	0.65
R-143a	0.90	1.34	0.18
R-12	1.06	1.18	0.49
R-22	0.99	0.84	0.40

R-600a	1.17	1.44	0.73

Table 8.18: Clusters of the Pure Component Effects

<i>ASHRAE #</i>	Property Clusters			<i>AUP</i>
	<i>C₁</i>	<i>C₂</i>	<i>C₃</i>	
R-32	0.596	0.388	0.017	1.53
R-152a	0.387	0.370	0.243	2.75
R-290	0.412	0.469	0.119	2.40
R-600	0.357	0.385	0.258	3.52
R-125	0.350	0.594	0.057	2.48
R-134a	0.351	0.425	0.225	2.88
R-143a	0.372	0.555	0.073	2.41
R-12	0.386	0.433	0.180	2.74
R-22	0.443	0.376	0.180	2.23
R-600a	0.349	0.431	0.220	3.35

Finally, the FR, MFR, pFR, and pMFR nondimensionalized property operator regions are calculated using Eq. 8.15 and shown in Table 8.19.

Table 8.19: The Nondimensionalized Property Operators for the Feasibility Region Types

Feasibility Region Type	Bounds	Ω_1	Ω_2	Ω_3
FR	<i>LL</i>	0.91	1.07	0.39
	<i>UL</i>	1.11	1.41	0.61
MFR	<i>LL</i>	0.87	0.60	0.15
	<i>UL</i>	1.26	1.46	0.91

pFR	<i>LL</i>	0.88	1.00	0.00
	<i>UL</i>	1.16	1.52	1.11
pMFR	<i>LL</i>	0.85	0.40	0.00
	<i>UL</i>	1.29	1.87	1.50

From these bounds, the cluster feasibility regions can be calculated using Eq. 6.7 resulting in Tables 8.20-8.23.

Table 8.20: The Property Cluster Bounds on the FR

<i>FR</i>	Property Clusters			<i>AUP</i>
	<i>C₁</i>	<i>C₂</i>	<i>C₃</i>	
Pt. 1	0.351	0.413	0.235	2.59
Pt. 2	0.310	0.482	0.207	2.94
Pt. 3	0.335	0.520	0.145	2.72
Pt. 4	0.381	0.485	0.367	2.92
Pt. 5	0.432	0.416	0.153	2.58
Pt. 6	0.398	0.384	0.218	2.79

Table 8.21: The Property Cluster Bounds on the MFR

<i>MFR</i>	Property Clusters			<i>AUP</i>
	<i>C₁</i>	<i>C₂</i>	<i>C₃</i>	
Pt. 1	0.367	0.251	0.382	2.37
Pt. 2	0.269	0.451	0.280	3.24
Pt. 3	0.352	0.589	0.059	2.48

Pt. 4	0.440	0.509	0.051	2.86
Pt. 5	0.629	0.298	0.073	2.00
Pt. 6	0.456	0.216	0.329	2.76

Table 8.22: The Property Cluster Bounds on the pFR

pFR	Property Clusters			AUP
	C_1	C_2	C_3	
Pt. 1	0.294	0.334	0.371	2.99
Pt. 2	0.251	0.433	0.316	3.51
Pt. 3	0.367	0.633	0.00	2.40
Pt. 4	0.432	0.568	0.00	2.67
Pt. 5	0.537	0.463	0.00	2.15
Pt. 6	0.354	0.306	0.340	3.26

Table 8.23: The Property Cluster Bounds on the pMFR

$pMFR$	Property Clusters			AUP
	C_1	C_2	C_3	
Pt. 1	0.310	0.144	0.546	2.74
Pt. 2	0.201	0.444	0.355	4.22
Pt. 3	0.312	0.688	0.00	2.72
Pt. 4	0.408	0.592	0.00	3.16
Pt. 5	0.766	0.234	0.00	1.69

Pt. 6	0.405	0.124	0.470	3.18

The resulting clusters of the FR, MFR, training set responses, and the training set pure component effects are plotted in Fig. 8.9. The clusters of the pFR, pMFR, training set responses, and the pure components are plotted in Fig. 8.10.

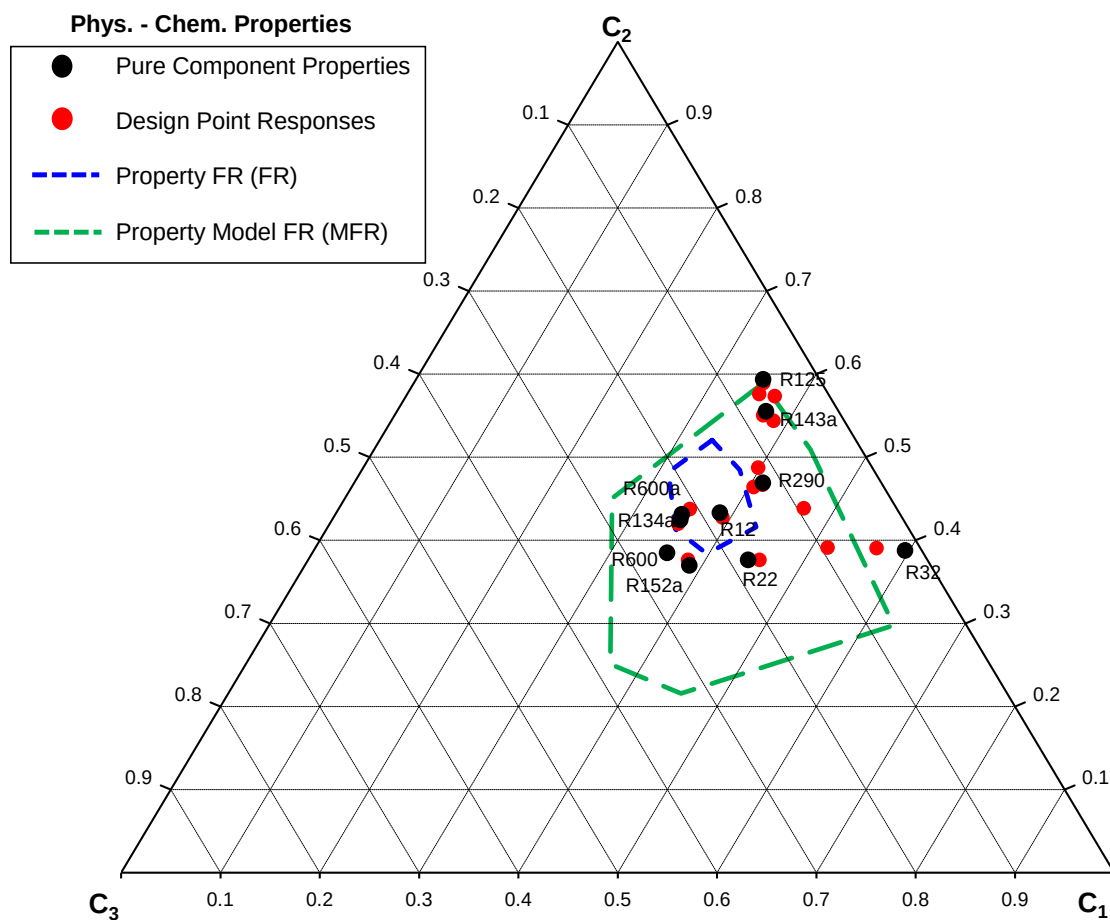


Figure 8.9: The Property Cluster Diagram Highlighting the Mixture Design, Pure Chemical Constituents, FR, and MFR.

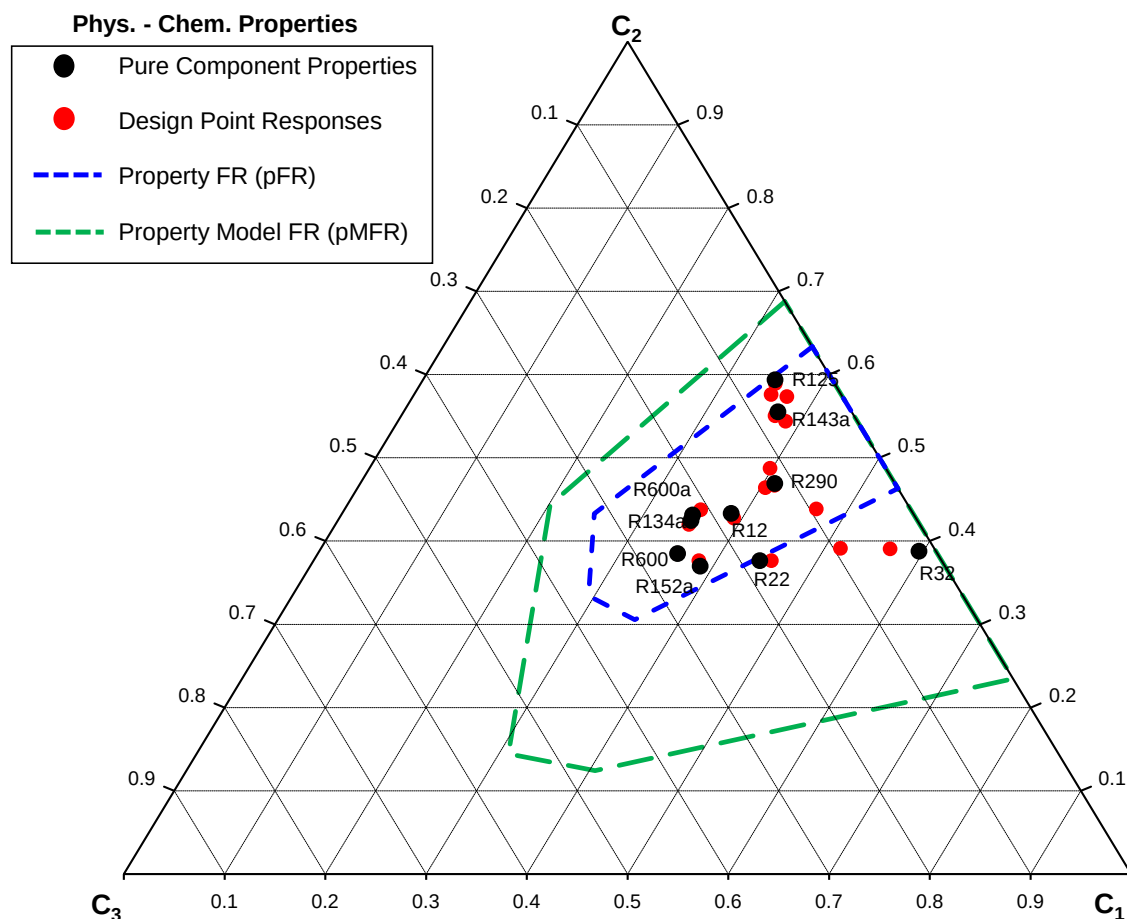


Figure 8.10: The Property Cluster Diagram Highlighting the Mixture Design, Pure Chemical Constituents, pFR, and pMFR.

As shown in both Fig. 8.9 and Fig. 8.10, the model feasibility region and target feasibility region completely overlap, meaning that all feasible mixtures estimated can be found by interpolation of the model. However, due to a poor fit of the the H_v model, there is sizable variability along C_3 , as evidenced by the variation of the shape of the FR vs. the pFR, and to a lesser extent the MFR vs. pMFR feasibility ranges. The larger of the domains (pFR and pMFR) are chosen to ensure that no feasible solutions are excluded from the design. Several refrigerants fall simultaneously within both the pFR and pMFR, but none of the candidates have AUP values that match the AUP values of the sink, with the exception of R-12. This result is expected since

R-134a was meant as a replacement for R-12. Since the potential environmental impact (*PEI*) of R-134a is 0.784/kg and the *PEI* of R-12 is 2.751/kg, R-12 will not meet the objective of lowering the *PEI* in this design problem. Furthermore, since no other mixture of components in the training set met the objectives of the design, it is desirable to either find new mixtures and/or design new refrigerants that can help meet the constraints. The remaining focus of this case study will be on the latter.

Due to large inventories of R-32 available to the consumer, it is desirable to find a way to use this refrigerant in the design. Since R-32 does not lie within the feasibility region, an additive must be used. The “additive” feasibility range must be bounded by the model feasibility range and the target feasibility range so that the resulting mixture meets both constraints simultaneously as shown in Fig. 8.11. To determine the combination of the two feasibility regions, we select the boundary point clusters of the appropriate pFR and pMFR and calculate the cluster at the pMFR intersection of a hypothetical component and R-32 that crosses the boundary of the pFR. The “additive domain” cluster boundary points are given in Table 8.24.

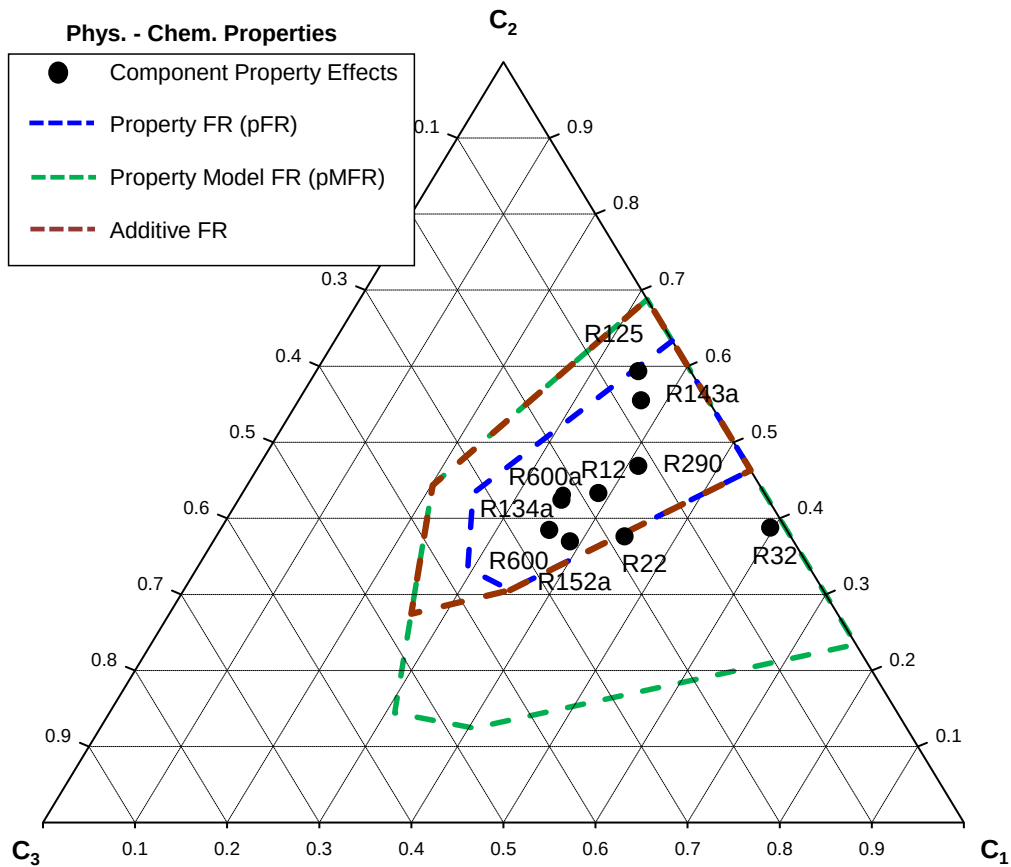


Figure 8.11: The Property Cluster Diagram Highlighting the Additive Feasibility Region.

The property operator upper and lower limit boundaries are then be calculated from Table 8.24 using minimization and maximization.

Table 8.24: The Additive Domain Feasibility Region Boundary Clusters and AUPs.

$pMFR$	Property Clusters			AUP
	C_1	C_2	C_3	
Pt. 1	0.269	0.594	0.138	3.26
Pt. 2	0.178	0.393	0.429	2.15

Pt. 3	0.258	0.120	0.622	2.67
Pt. 4	0.322	0.110	0.568	2.72
Pt. 5	0.583	0.199	0.218	4.22
Pt. 6	0.335	0.540	0.125	3.23

Table 8.25: The Additive Domain Feasibility Region Bounds.

Additive FR	Ω_1	Ω_2	Ω_3	AUP*
<i>LL</i>	0.85	0.59	0.00	2.15
<i>UL</i>	1.16	1.87	1.50	4.22

* Unbalanced AUP from Table 8.24.

It should be noted that since the additive region is an amalgam of the pFR, pMFR, and a hypothetical component mixed with R-32, it is no longer balanced. In other words, a recreation of the “additive” feasibility region using the values on Table 8.25 and the feasibility construction method listed in Chapter 6 would result in an overestimation of the domain. In this situation, it is valid to use the domain boundaries listed in Table 8.24 to test Rules 1 and 6, while using the unbalanced minimum and maximum *AUP* values to test Rules 2 and 7. Rules 3 and 8 must be satisfied by matching the *AUP* values of the pFR, pMFR, and potential candidate solution or mixture. For the solutions estimated using the forward approach, the unbalanced solution domain (i.e. Table 8.24) provides an additional test for potential candidate solutions that meet the constraints listed in Table 8.25.

After the target additive region has been identified, compounds that can satisfy it are identified using the aCAMD method. Since the additive is a refrigerant component, its desirable for the molecular weight not be too high and the search is limited to compounds with molecular

weight between 30 and 250 g/mol. The group contribution expressions for the target properties become as shown:

$$H_v = H_{v0} + \sum_{g=1}^{N_g} n_g h_{v1} + \omega \sum_{s=1}^{N_s} n_s h_{v2} \quad (8.46)$$

$$T_c = T_{c0} \cdot \ln \sum_{g=1}^{N_g} n_g t_{c1} + \omega \sum_{s=1}^{N_s} n_s t_{c2} \quad (8.47)$$

$$P_c = P_{c1} + \left(P_{c2} + \sum_{g=1}^{N_g} n_g p_{c1} + \sum_{s=1}^{N_s} n_s p_{c2} \right)^{-2} \quad (8.48)$$

$$M = \sum_{g=1}^{N_g} n_g m_0 \quad (8.49)$$

The property operators for each property can be formed according to Section 6.3 and Section 8.2. All group contribution property values are taken from Marrero and Gani (2001). In this design, ten different acyclic molecular groups are considered for generating the potential candidates. These groups are typically found in refrigerants and often display less environmental impact. The property ranges in terms of normalized operators shown in Tables 8.25 are augmented to reflect the molecular weight constraints (e.g. $0.3 \leq \Omega_4 \leq 2.5$) using a reference value of 100. The first order group contribution parameters are as follows.

Table 8.26: Property Data of Selected First Order Groups (Marrero and Gani 2001)

Group g_1	Name	FBN	t_{c1}	p_{c1}	h_{v1}	m_0
1	CH ₃	1	1.751	0.018615	0.217	15
2	CH ₂	2	1.3327	0.013547	4.910	14

3	CH	3	0.5960	0.007259	7.962	13
4	CH ₂ F	1	3.3179	0.023315	8.238	33
5	CF	3	2.1633	-0.010120	6.739	31
6	CF ₂	2	0.8543	0.018572	1.621	50
7	CF ₃	1	1.7737	0.048565	7.352	69
8	CH ₂ O	2	2.4217	0.017954	9.997	30
9	CH ₃ O	1	3.4393	0.020084	5.783	31
10	CH ₂ =CH	1	3.2295	0.025745	4.031	27

It should be noted that although some of these groups may be represented by smaller, first order groups (e.g. CF may also be represented by a C and F group), in practice, better model agreement is obtained by representing the arrangement by the group with the largest molecular weight (Marrero and Gani 2001). In terms of first order groups, the parameters are rewritten as non-dimensionalized property operators and *AUP* values as shown in Table 8.27. For ease of comparison, the molecular weight property operator is not included in *AUP* of the mixture. It will serve as an additional screening step for the design molecules.

Table 8.27: Non-dimensional Property Operators and AUPs of Selected Molecular Fragments

Group g_1	$\Omega_{1g_1}^M$	$\Omega_{2g_1}^M$	$\Omega_{3g_1}^M$	$AUP_{g_1}^M$
1	0.35	0.37	0.02	0.74
2	0.27	0.27	0.49	1.03
3	0.12	0.15	0.80	1.06
4	0.66	0.47	0.82	1.95
5	0.43	-0.20	0.67	0.90

6	0.17	0.37	0.16	0.70
7	0.35	0.97	0.74	2.06
8	0.48	0.36	1.00	1.84
9	0.69	0.40	0.58	1.67
10	0.65	0.51	0.40	1.56

Using Eq. 6.43, an inequality expression for each property is generated. Using only first order group property operators, the maximum possible number of each group is calculated and shown in Table 8.28.

Table 8.28: Non-dimensional Property Operators and AUPs of Selected Molecular Fragments

Group g_1	Formula	Calculated n_{g1}^U	Buffered n_{g1}^U
1	CH ₃	5	6
2	CH ₂	4	4
3	CH	3	4
4	CH ₂ F	2	2
5	CF	4	4
6	CF ₂	5	5
7	CF ₃	2	2
8	CH ₂ O	2	2
9	CH ₃ O	2	2
10	CH ₂ =CH	2	3

However, as noted in Section 8.3, these values may not be the true limits of the number of groups because of the potential for negative second order property operators. Also, since some

of the first order groups listed in Table 8.27 contain smaller first order groups (i.e. Table 8.27 is not a complete set of first order groups in the system), second order group instances containing the smaller first order groups must also be accounted for. The second order groups with known contributions that can be built from the first order groups listed in Table 8.27 are given in Table 8.29 (Constaninou and Gani 1994). Those groups containing smaller first order groups (i.e. C, F, etc.) are denoted accordingly.

Table 8.29: Properties of 2nd Order Groups built from 1st Order Groups (Marrero and Gani, 2001)

Group g_2	Formula	t_{c1}	p_{c1}	h_{v1}
1	$(CH_3)_2CH$	-0.0471	0.000473	-0.419
2	$CH(CH_3)CH(CH_3)$	0.5602	-0.003207	0.532
3	$CH_2=CH-CH=CH_2$	0.4214	0.000792	1.632
4	$CH_3-CH=CH_2$	-0.0172	-0.000101	0.064
5	$CH_2-CH=CH_2$	0.0262	-0.000815	-0.060
6	$CH-CH=CH_2$	-0.1526	-0.000163	0.004
7	$CH_2-O-CH=CH_2$	0.2900	-0.000432	0.372
8	$CH_3-O-CH=CH_2$	0.2900	-0.000432	0.372
9*	$(CH_3)_3C$	-0.1778	0.000340	-0.417
10*	$CH(CH_3)C(CH_3)_2$	0.8994	-0.008733	0.623
11*	$C(CH_3)_2C(CH_3)_2$	1.5535	-0.016852	5.086
12*	$C-CH=CH_2$	-0.1526	-0.000163	0.004

* 2nd order groups built from smaller first order groups

The 2nd order property contributions can then be expressed in terms of non-dimensionalized property operators and the augmented property index as shown in Table 8.30.

Table 8.30: The non-dimensionalized property operators of 2nd order groups

Group g_2	$\Omega_{1g_2}^M$	$\Omega_{2g_2}^M$	$\Omega_{3g_2}^M$	$AUP_{g_2}^M$
1	-0.0094	0.0095	-0.0419	-0.042
2	0.1120	-0.0641	0.0532	0.101
3	0.0843	0.0158	0.1632	0.263
4	-0.0034	-0.0020	0.0064	0.001
5	0.0052	-0.0163	-0.0060	-0.017
6	-0.0305	-0.0033	0.0004	-0.033
7	0.0580	-0.0086	0.0372	0.087
8	0.0580	-0.0086	0.0372	0.087
9	-0.0356	0.0068	-0.0417	-0.070
10	0.1799	-0.1747	0.0623	0.068
11	0.3107	-0.3370	0.5086	0.482
12	-0.0305	-0.0033	0.0004	-0.033

The 2nd order groups with negative AUP^M (i.e. groups 1, 5, 6, and 12) contain the 1st order groups CH₃, CH₂, CH, and CH₂=CH, meaning that a molecule containing these second order groups could theoretically contain a higher number of the listed 1st order groups than shown in Table 8.27. To protect against this scenario, the maximum number of 1st order groups are buffered by 20% where necessary, resulting in the buffered column in Table 8.27. A buffer of 20% was chosen because second order groups generally correct first order groups by $\pm 10\%$.

Proceeding with the additive design using the aCAMD, all possible combinations of first order and second order groups are then estimated subject to the constraints placed on the search space shown in Table 8.25. In addition, the customer has indicated a desire to limit the number

of dissimilar groups to two. Following the procedure outlined in Table 8.3 and Fig. 8.4-8.6, potential candidate molecules are generated. The clusters of the first order groups used in the aCAMD are given in Table 8.31.

Table 8.31: First Order Group Clusters

Group g_1	$C_{1g_1}^M$	$C_{2g_1}^M$	$C_{3g_1}^M$
1	0.471	0.500	0.029
2	0.259	0.263	0.477
3	0.112	0.137	0.751
4	0.340	0.239	0.422
5	0.479	-0.224	0.745
6	0.243	0.527	0.230
7	0.172	0.471	0.357
8	0.263	0.195	0.542
9	0.412	0.241	0.347
10	0.413	0.329	0.258

As shown in Fig. 8.12, one group, CF, resides outside the positive cluster domain. Since, any mixtures containing CF would need to reside within the positive cluster domain, any “pure group” mixtures of this group are automatically discarded. Of the remaining groups, only two “pure groups”, CF₂ and CH₃, are capable of producing candidates that meet Rule 1. *AUP* values for each of the candidates can then be tested and matched against the *AUP* range of the additive feasibility region. Between 4 and 5 CF₂ groups are capable of meeting this constraint and the maximum number of groups constraint. The number of CH₃ groups meeting these constraints are between of 3 and 5. Calculating the Free Bond Number (*FBN*) of these molecules for non-ring

compounds indicates that neither set contains a thermodynamically stable molecule and all pure groups are discarded.

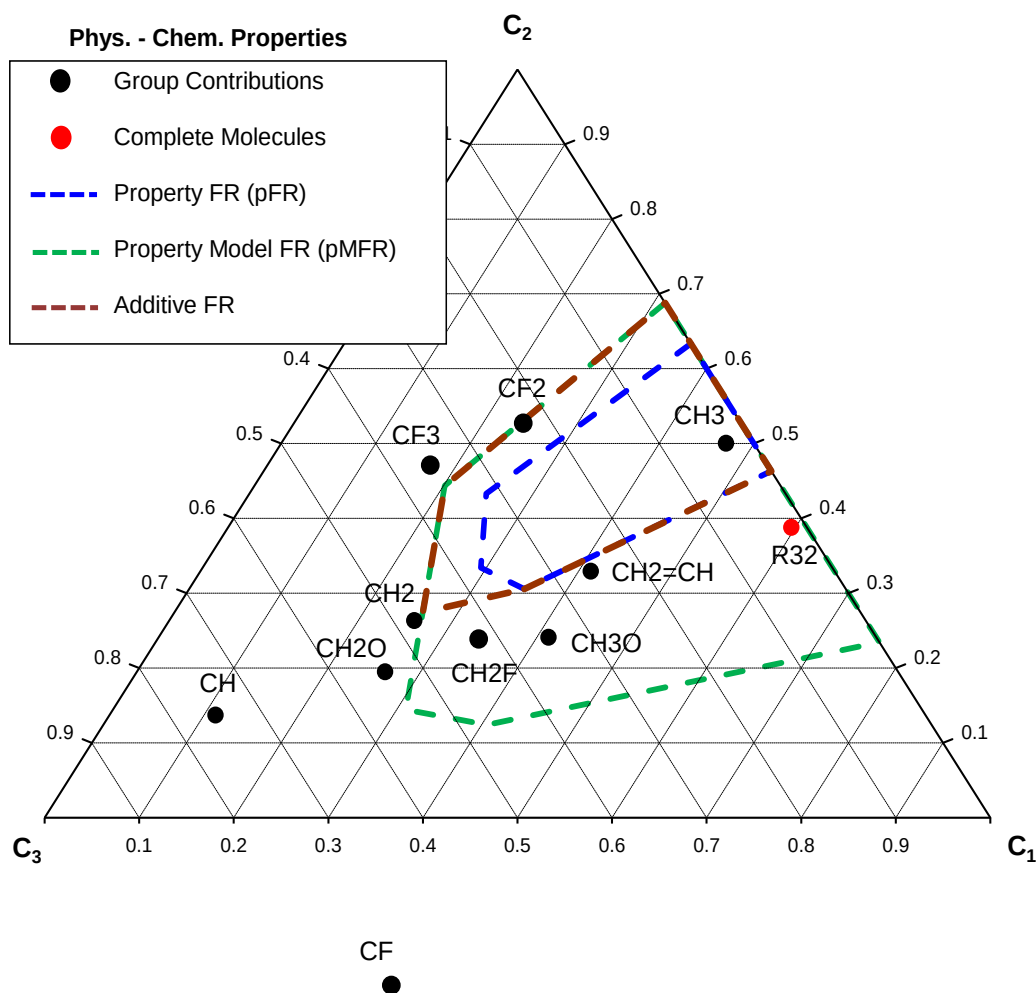


Figure 8.12: 1st Order Group contributions in the Cluster Domain

For molecules having two dissimilar groups of Table 8.26, 389 potential non-ring candidate molecules can be built. Using the procedure outlined in Fig. 8.6, each non-ring candidate molecule was tested against Rule 4 (i.e. $FBN = 0$). In total, 321 molecules failed Rule 4 and were discarded. Five more molecules failed Rule 14 and were discarded leaving 63 potential candidate molecules. Finally, 56 molecules failed to fall within the “additive” feasibility region

constraints of Table 8.24, leaving 7 potential candidate molecules. Two of the seven molecules, R-161 and Methoxymethane, fall just outside the cluster domain because it is unbalanced. The remaining five potential candidate solutions are shown in Table 8.32 and Fig. 8.13.

Table 8.32: Candidate Molecules

ASHRAE #	Name	Formula	Ω_1^M	Ω_2^M	Ω_3^M
R-290	Propane	CH ₃ -CH ₂ -CH ₃	0.967	1.016	0.534
R-272ca	2,2-Difluoropropane	CH ₃ -CF ₂ -CH ₃	0.871	1.116	0.206
-	2,2,3,3-Tetrafluorobutane	CH ₃ -(CF ₂) ₂ -CH ₃	1.042	1.487	0.368
R-1243	3,3,3-Trifluoropropene	CH ₂ =CH-CF ₃	1.001	1.486	1.138
R-143m	Methyl trifluoromethyl ether	CH ₃ O-CF ₃	1.043	1.373	1.314

It should be noted that only these 5 candidate molecules were fully constructed by the algorithm. All five of the candidate molecules meet the property constraints on critical temperature, critical pressure, enthalpy of vaporization, and molecular weight of the molecule. Of them, R-290 was part of the training set and was discarded because its attribute values did not fall within the consumer targets. The difference between the designed molecule and training set molecule can be attributed to two sources of error: (1) the widening of the target domain because of a poor fit in Eq. 8.45 and (2) a systematic underestimation of the enthalpy of vaporization by the Marrero and Gani GCM method (2001). The first source of error could be improved by developing a stronger relationship for Eq. 8.45. However, since the aCAMD is built to only exclude infeasible solutions and return candidate solutions, with no promise as to the suitability of the designed molecules, the effect of this uncertainty is limited.

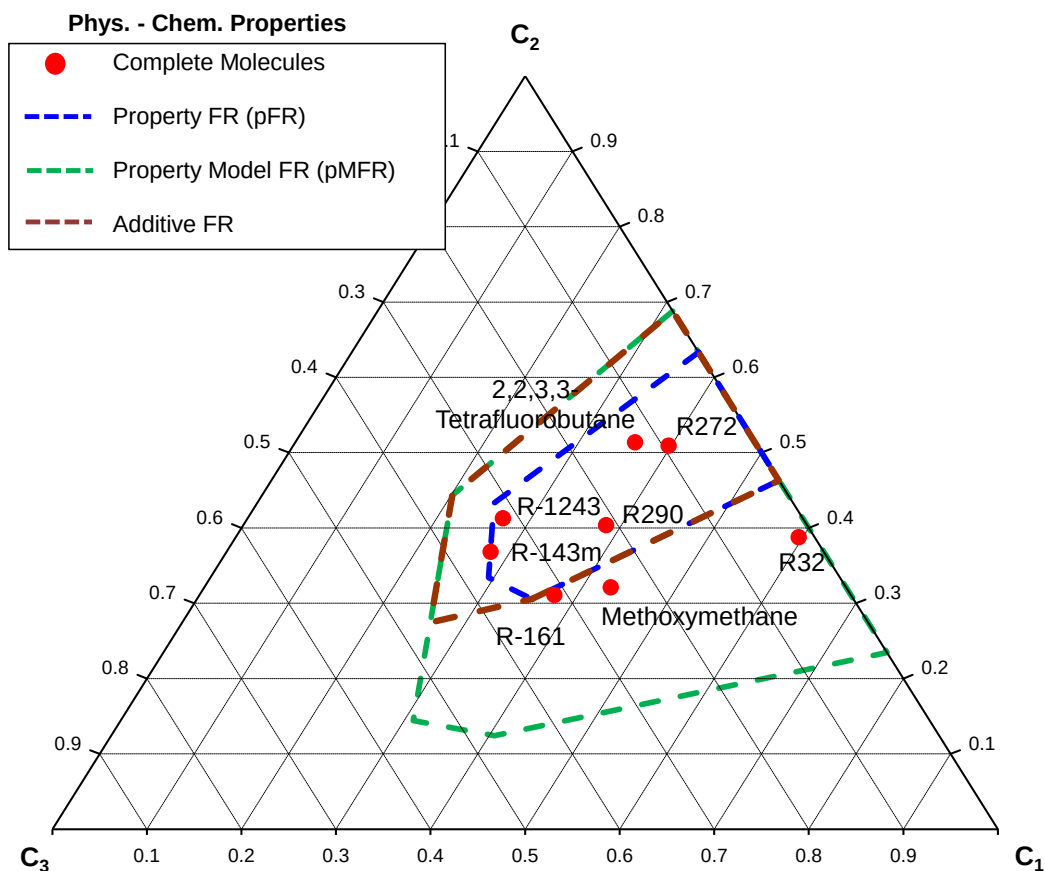


Figure 8.13: Candidate Molecules in the Physical-Chemical Property Cluster Domain

For the second source of error, plots of the enthalpy predicted vs. actual responses of the enthalpy of vaporization shown in Fig. 8.14a indicate a severe incongruity in the parameterization of the terms by Marrero and Gani (2001).

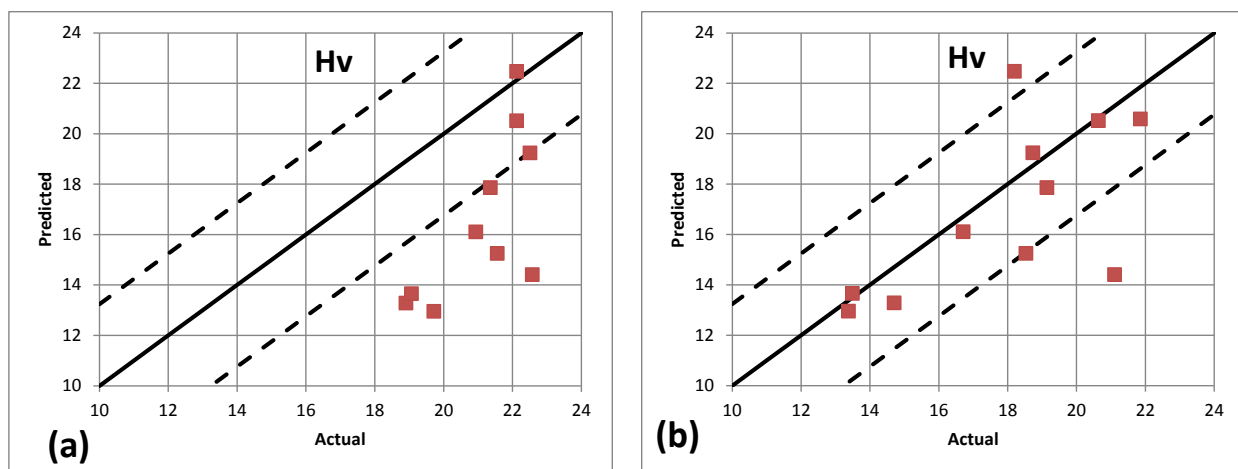


Figure 8.14: Predicted vs. Actual Latent Heat of Vaporization using: (a) GCM at 298K and (b) GCM at the Boiling Point T_b .

Upon inspection, it appears that the Marrero and Gani GCM prediction is actually providing the enthalpy of vaporization at the boiling point and not at the standard state of 298K and 1 bar as was reported. Using the correlation in Eq. 8.50 developed by Watson (1943), the predicted enthalpy of vaporization is corrected to the standard state and plotted in Fig. 8.14b.

$$H_v = H_{vb} \left(\frac{1 - \frac{298.15}{T_c}}{1 - \frac{T_b}{T_c}} \right)^{0.38} \quad (8.50)$$

Although the prediction of the enthalpy of vaporization improves, a significant amount of error over the approximately ± 1.61 standard deviation reported by Marrero and Gani (2001) remains in the system. The error is primarily a result of the contributions of the halide groups like CF_2 and CF_3 which also affect the critical temperature (T_c) as shown in Fig. 8.15a. The critical pressures (P_c) in Fig. 8.15b did not appear to be affected by these groups.

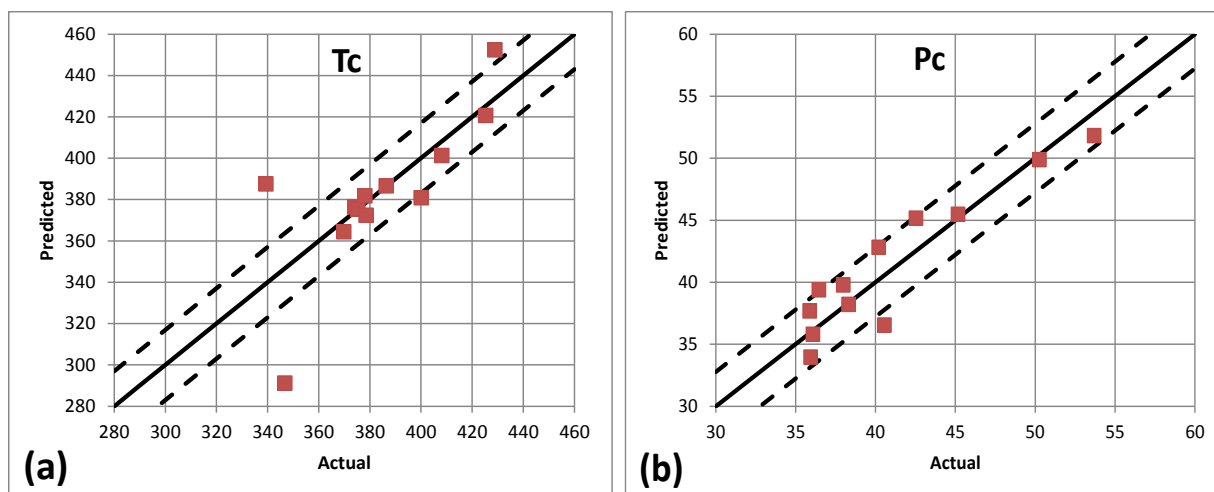


Figure 8.15: Predicted vs. Actual Critical Properties: (a) Temperature and (b) Pressure.

These results suggest it may be beneficial to reparameterize the GCM method for only the groups of those molecules in the training set since they are used to build the candidate molecules. This type of approach is discussed in more detail in Chapter 9.

Of the four remaining candidate molecules, the *PEI* value provided by the WAR GUI Version 1.0.17 (Young and Cabezas 2000) could only be determined for a single component, R-1243. It was at 6.068 *PEI*/kg which is significantly higher than R-134a at 0.3255 and R-32 is thus ruled out as a potential candidate. Further environmental study of the 3 species without a *PEI* is warranted but is beyond the scope of this dissertation.

Based on these results, a binary mixture involving R-32 could not be found to satisfy the constraints of the design. However, noting that the *PEI* of R-161 is at 0.00406 and that it is close to the feasibility region, it is quite possible that a ternary mixture involving R-32 and R-161 could be utilized. To determine which of the tested components to add, it is necessary to evaluate the difference between the *AUP* of the target sink and the *AUP* of the R-32/R-161 mixture. It is found that the critical pressure cluster, represented by the property cluster C_2 , is too low and needs to be raised. Of the components utilized in this design, R-125 has the lowest

P_c value, which corresponds to the highest C_2 cluster value, which should help to balance the binary mixture. However, R-125 also has a high PEI value of 2.158. Making the assumption of linear mixture behavior and using lever-arm analysis, a ternary mixture of R-32/R-125/R-161 of 33/10/57 wt.% is found to simultaneously match both the AUP of the sink, cluster feasibility region, and PEI of 0.3255, matching R-134a. Subsequent reduction of R-125 will lower the PEI but at a cost of lowering the amount of R-32 that can be utilized in the design. In a similar experimental study comparing the cycle performance of R-32/R-125/ R-161 against R-407c, Han *et al.* (2007) found a mixture of 23/25/52 wt% to be sufficient for replacement for R-407c. Considering that R-407c is a 15/34/51 wt% mixture of R-32, R-125, and R-134a, it is reasonable to conclude that this mixture is similar to R-134a and can be used to corroborate these results. Accordingly, it is found that this method predicts a R-407c mixture between 15/34/51 wt% and 32/39/29 wt% which nearly encompasses the result from Han *et al.* (2007). The major difference between the two results is the amount of R-125 in the mixture and is a result of the experimental solution not being bound by the PEI , as was done in this case study. Had this constraint been relaxed, the R-125 fractions would have been severely reduced. There is also some indication in the gradients of the mixture responses provided by Han *et al.* (2007) that the mixing may be nonlinear. Further experimentation on this candidate solution is warranted.

In summary, the overall confidence that this solution is the global solution to the design problem is calculated by applying Eq. 8.13 to the property-attribute relationships in Table 8.9 and the GCM models in Marrero and Gani (2001), resulting in 88.4% and 98.1% for each step. However as noted earlier, significant errors in the calculation of H_v by Marrero and Gani (2001) suggest that the error in the second step is much larger and could potentially reduce the confidence by another 20-40%.

8.5. Summary

In conclusion, an attribute computer aided molecular design (aCAMD) technique using property clusters and the reverse problem formulation was developed and discussed in this chapter. The method allows for the complete and efficient solution of product design problems that rely on regression models to describe consumer attributes and property models based on group contribution. The use of smaller building blocks based on functional groups allows the technique to construct molecules that are not part of the original experimental data set, a significant contribution to the product design problem. In addition, because of the uncertainty encountered in defining consumer attributes, the method uses relaxations of the target solution domains in order to ensure that only infeasible candidate solutions are excluded from the design. However, the efficiency of the method suffers in situations of high uncertainty that are often found in the multiscale design of chemical products. In the case study, the accuracy (and efficiency) of the design suffered from two main sources of uncertainty: (1) data driven models in property-attribute relationship and (2) poorly fit group contributions in the property-molecular architecture relationship. For case 1, better models may be obtained by reconfiguring the problem in terms of underlying latent variables using decomposition techniques. Decomposition is a mathematical technique that can determine the most valuable underlying latent variables. Latent variables are, by definition, mathematical inferences of underlying independent variables that control the response of the system. The ability of decomposition to handle large sets of multivariate data suggests it could be utilized to evaluate the structural information of a chemical product over multiple scales. The use of decomposition and latent variable models to design chemical products at multiple scales is discussed in Chapter 9.

For case 2, it has been shown that it may be beneficial to parameterize the property-molecular architecture relationship on a case-specific basis in order to improve predictive power. This is especially true for chemical products with architecture that exists over multiple scales. One of the most common and efficient ways to describe this architecture is to use characterization to generate enormous multivariate data sets and derive case specific combinatorial techniques (based on GCM) according to the type of characterization used. The use of characterization based group contribution (cGCM) is discussed in more detail in Chapter 9.

Chapter 9

Characterization based Computer Aided Molecular Design (cCAMD)

As discussed throughout this dissertation, combinatorial techniques like GCM are valuable methods for efficiently designing molecules for specific end uses (Constantinou and Gani 1994; Marrero and Gani 2001; Gani 2004; Hill 2009). Unfortunately, in chemical product design, the associated properties of concern are most often consumer attributes which do not have group contribution parameters (Hill 2004; Hill 2009). This limitation was partially addressed in Chapter 8 by mapping the consumer attribute information to a set of intensive properties at the macroscale that can be described by group contribution (Solvason *et al.* 2009). This step is often performed via chemometrics which defines an empirical relationship through the use of design-of-experiments (DOE) and multivariate-linear-regression (MLR). Any uncertainty in the relationship between the attributes and properties is handled by increasing the size of the feasibility regions in the property domain (Solvason *et al.* 2009). However, if the attribute-property relationship is poorly defined, then the feasibility regions are enlarged such that any RPF efficiency gains are lost and the overall accuracy of the design is reduced. For multiscale chemical product design this observation is especially apparent. For example, the consumer attribute “softness” in facial tissue is a function of molecular scale structure (e.g. backbone flexibility), microscale structure (e.g. fiber coarseness), and mesoscale structure (e.g. fiber orientation) of the product. Using only GCM (i.e. molecular scale) properties to build the attribute-property relationship would fail to capture the architecture effects of the other length

scales on the product's attributes, resulting in a poor attribute-property relationship, a broad feasibility region, and many, poorly designed chemical product candidates that may or may not be true solutions. In contrast, if the system's attributes and properties can be described by only molecular scale architecture (e.g. solvents), then the system can be completely modeled by GCM (Eljack *et al.* 2007). It is for the former, multiscale system, that a new method called characterization based computer aided molecular design (cCAMD) has been developed in Chapter 9.

The cCAMD method uses characterization data to describe the molecular architecture of the system, decomposition techniques to consolidate that information into a linearized latent property subdomain, and chemometric techniques to relate the latent property subdomain to the attribute domain. With this solution structure, attribute target data can be converted into latent property targets where molecular architectures can be linearly combined to deliver them using the property clustering algorithm. This approach differs from the multivariate mixture design approaches discussed in Chapter 3 in the type of building blocks used in the design. Whereas Muteki and McGregor (2006) and Garcia-Serna *et al.* (2011) search large databases of molecules for pure components or combine pure components to generate candidate mixtures, this method uses a much smaller set of molecular groups present in the original training set to build candidate molecules and candidate mixtures. The result is a method that generates a complete set of candidate molecules, void of database completeness limitations.

For this method to work properly, important multi-scale molecular architecture features must be appropriately captured by various characterization techniques and converted into molecular group-like features which can be recombined in a characterization based group contribution method (cGCM) to generate new chemical products. Since infrared (IR) and near

infrared (NIR) spectroscopy techniques are particularly adept at capturing molecular scale and microscale features of chemical products, they were chosen to highlight the method (see Section 9.1.1).

Once the spectral features have been captured, decomposition techniques are applied to determine the latent property subdomain in order to improve the combinatorial efficiency and certainty of the designed candidate molecules. The use of latent property descriptors requires the construction of an attribute-latent property relationship in order to provide a vehicle for mapping attribute targets into the latent subdomain. The decomposition and relationship construction techniques for the latent domain are discussed in Section 9.1.2.

In Section 9.2 the necessary changes to the property clustering method to handle latent property descriptors of molecular architecture and targets are examined. In particular, the design must be augmented to handle negative mixture fractions. The section concludes with the cluster conversion algorithm for cCAMD.

In Section 9.3 the candidate generation algorithm for cCAMD is developed and discussed. Changes to the structure generation approach and additional structure checks are highlighted. The resulting algorithm is capable of simultaneously building candidate molecules with multiple types of molecular architecture that exist across multiple scales.

In Section 9.4 a case study on the design of a pharmaceutical excipient formulation for direct compressed acetaminophen tablets is used to illustrate the method. Two types of designs are conducted, a mixture design to highlight the impact of the property clustering technique, and a molecular design to highlight the newly developed characterization based group contribution method (cGCM). The chapter concludes with a summary in Section 9.5.

9.1. Integrating Multivariate Data & Models with the RPF

In general, the objective of the cCAMD method is to enumerate all possible molecules and molecular architectures that meet a set of target attributes by using a mapping function to move the chemical product design problem into sub-domain with linear programming characteristics. Identifying a proper sub-domain in which to conduct the design is paramount to ensuring a repeatable, accurate, and physically meaningful solution occurs. A suitable sub-domain should possess the following qualities: (1) the sub-domain latent property-molecular architecture relationship should be linear to ensure computational efficiency and solution global optimality, (2) the sub-domain should be capable of representing all relevant molecular architecture important to the design, and (3) sub-domain mapping noise should be minimized by using latent property-attribute relationships with strong predictive power (Solvason *et al.* 2009, Solvason *et al.* 2010).

Many techniques exist that can be used to find a suitable property sub-domain. One of the most common is the decomposition of property and molecular architecture information into an eigenfunction. When the technique minimizes the mahalanobis distance to build the underlying, orthogonal latent variable structure, it is called principal component analysis (PCA) (Johnson and Wichern 2007). Alternatively, when information about the relationships between molecular architecture and the attributes of the system is available, network component analysis (NCA) can be used to find the underlying sub-domain using a MINLP formulation proposed by (Tolle *et al.* 2009). It should be noted that in order for either of these decomposition techniques to be effective, large amounts of heavily correlated data that describe the molecular architecture will be required. When this requirement is combined with the second criteria of a complete description of the system, it is clear that data-gathering techniques spanning multiple scales will

need to be utilized. It has been shown that characterization tools and techniques are excellent methods of generating data and information about the system.

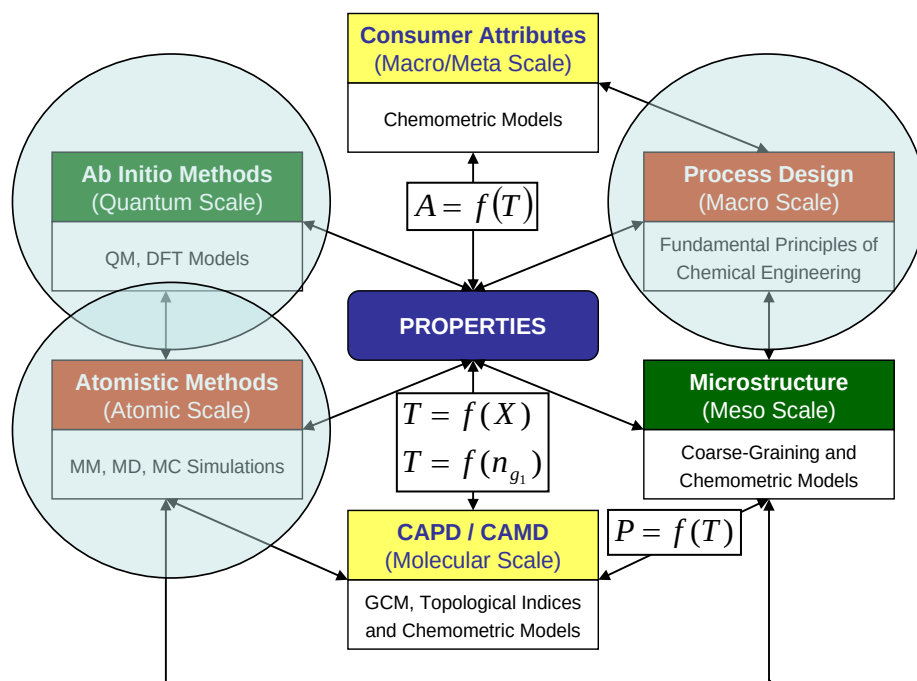


Figure 9.1: A Multiscale Reverse Problem Formulation. The Scales Not Blocked Out are Capable of being Described using the Techniques Developed in this Section.

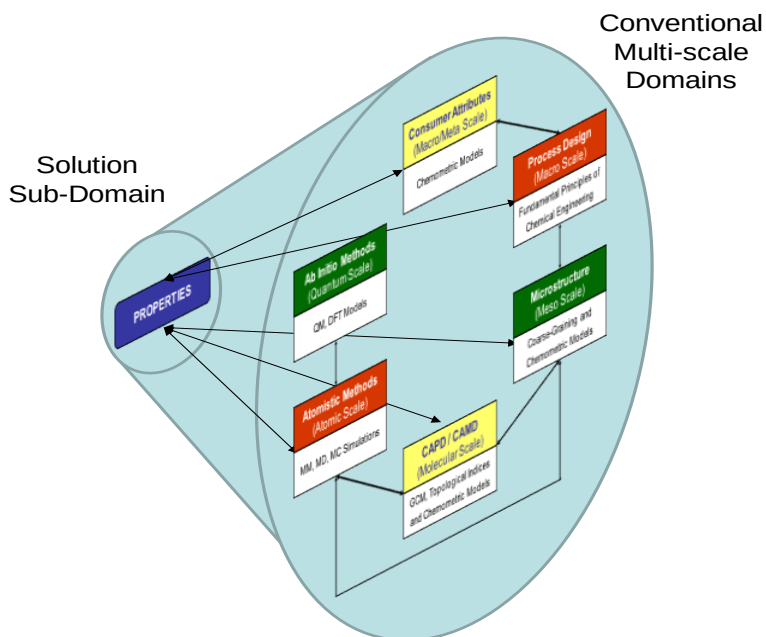


Figure 9.2: A Cross-section of the Reverse Problem Formulation Applied to a Multiscale Chemical Product Design.

As shown in Fig. 9.1 and 9.2, most chemical products with multiscale molecular architecture are usually present in a solid form and, in particular, rely on microstructure at the micro- through meso-scales as well as molecular structure at the molecular scale to deliver the needed consumer attributes. Unlike the aCAMD method of Chapter 8 which used only molecular scale information, the cCAMD method uses characterization property data (P) of an experimental training set to capture molecular architecture at multiple scales followed by a decomposition to find a underlying latent property domain (T).

$$P = f(T) \quad (9.1)$$

This type of approach overcomes the limitation of the aCAMD method by capturing a complete description of the relevant molecular architecture important to the product. Likewise, documented characterization responses for specific types of molecular architectures can also be decomposed to find a set of latent domain parameters to be used in a cGCM or linear mixture design:

$$T = f(n_{g1}) \quad (9.2)$$

$$T = f(X) \quad (9.3)$$

where n_{g1} are the number of distinct first order characterization based groups, g_1 , and X is the vector of pure component mass fractions in the product. Finally, the targets for the design are obtained by developing the attribute-latent property relationship and mapping the consumer attributes (A) into the latent domain.

$$A = f(T) \quad (9.4)$$

With the targets defined, characterization based groups and pure component molecular architectures can be mixed to determine the chemical product candidates. The strength of the

method is predicated on two main aspects, (1) the proper characterization of the system and (2) the proper decomposition to find a suitable latent variable domain.

9.1.1. Characterizing the System with IR/NIR Spectroscopy

As reviewed in Chapter 3, infrared (IR) and near infrared (NIR) spectroscopy can be useful characterization techniques for the determination of chemical constituency, molecular structure, and nano-, micro-, and meso-scale orientation and alignment. The characterizations are often performed on a training set of molecules defined by an experimental design used to explore the interesting facets of a product's design. Recent developments in these spectroscopic techniques have been shown to provide better determination of overtone and combination bands of familiar molecules and improved descriptions of bonding phenomena such as van der Waals forces (Nesbitt 1988; Saykally 1989; Hutson 1990; Heaven 1992; Iachello and Levine 1995), making them invaluable for the product designer.

Since, as discussed in Chapter 8, significant computational efficiency is often gained when the design problem is approached using group theory, it is beneficial to relate the IR/NIR characterizations to functional groups. Using the concept of symmetry, many atoms and combinations of atoms in a symmetrical molecule can be considered to be in the same chemical environment, and by extension, the vibrational motions and absorptions or IR and NIR spectra will be identical. Hence, it makes more sense to parameterize specific group vibrational spectra as opposed to atomic or complete molecular spectra as was done in Chapter 3. Groups can be considered as the formation of atoms interacting in a molecular association. The calculation of group frequencies is a summation of vibrational motions described by harmonic oscillation, anharmonic oscillation, first order coupling, and Fermi resonance (second order coupling). Various concepts regarding rotational axes, reflection planes, inversion centers, improper

rotational axes, and point groups can be used to mathematically justify this concept, but are beyond the scope of this dissertation (Workman and Weyer 2008). Instead, a brief summary of the vibrational frequencies in terms of normal coordinates in group theory is presented.

Many factors are responsible for determining the frequency of absorbance for a specific molecular architecture. The strongest fundamental bands correlate with the strongest bonds and inversely correlate with the mass of an atom (Socrates 2001). Overtone bands arise from anharmonicity of the fundamental bands and can be used to differentiate between group bonds that absorb in the same frequency. However, combination bands and coupling may cause some of these vibrations to shift, degenerate, or become inactive in IR or NIR. The overlap of features will need to be addressed by more sophisticated approaches (Stephenson *et al.* 2001). One such technique is to utilize the group contribution method (GCM).

As discussed in Chapters 4, 6, and 8, GCM is a property prediction technique based on the UNIFAC method of summing individual property contributions from specific functional groups in a molecule. 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 method and help to correct for polyfunctional compounds with more than four carbon atoms in the main chain (Marrero and Gani 2001). The complete GCM expression is as follows:

$$P^M = \overbrace{\sum_{g_1} n_{g_1} P_{g_1}}^{1^{\text{st}} \text{ Order}} + w_S \overbrace{\sum_{g_2} n_{g_2} P_{g_2}}^{2^{\text{nd}} \text{ Order}} + w_T \overbrace{\sum_{g_3} n_{g_3} P_{g_3}}^{3^{\text{rd}} \text{ Order}} \quad (4.15)$$

The method proposed in this chapter uses characterization data from IR and NIR spectroscopy to define the group contributions in the GCM expression in order to predict the

spectra of a full molecule. Additionally, molecular architecture information beyond the scope of traditional GCM, such as group orientation and crystal polymorphism, will be included as long as it can be ascertained from IR and NIR spectroscopy. In order to handle these special constraints, a modified characterization based group contribution is proposed that parameterizes the groups in a nested structure. To be effective, the dissemination of characterization data must follow a set of rules designed to shadow those developed by Marrero and Gani (2001). First, the characterization must be able to completely quantify each individual molecular group used in the design. Within a group, the number of absorbance frequencies can be calculated by modifying Eq. 3.27 to account for bonds by assuming each bond is attached to a hydrogen atom.

Rule 17. The total number of fundamental vibrations of a group will be calculated using Eq. 3.27, where each open ended bond is estimated as a hydrogen atom.

This assumption will grossly overestimate the number of fundamental modes of vibration, but will serve as a necessary constraint. For instance, this method will estimate that a CH₂ group had 9 fundamental modes, when, in fact, it only has 6 fundamental modes of vibration. It should be noted that this overestimation is only a concern when molecular groups have a small number of atoms with strong absorbances (e.g. C=O). For larger groups, significant degeneracy from interactions leads many fundamental tones to become inactive. For these situations, the inactive tones can be ignored; the use of only the physically active tones listed in various texts will be sufficient.

Rule 18. The number of fundamental tones calculated by Rule 17 will be reduced to only the tones that have been experimentally validated. No quantum mechanical estimations of tones will be utilized.

For instance, the physically identified fundamental bands of CH₂ are presented in Table 9.1 (Socrates 2001; McKenzie 2002). The intensity of absorption of each fundamental absorption is listed in terms of a relative rates of strong (s), medium (m), medium-strong (m-s), and inactive (i). Although it was expected that the CH₂ group would have 9 fundamental tones, only 3 could be experimentally verified: symmetrical stretching, asymmetrical stretching, and scissoring bend. The remaining bands were found to be *i* inactive.

Table 9.1: Fundamental Vibrations of the CH₂ Group (Socrates 2001; McKenzie 2002)

Band	Wavelength Region (cm ⁻¹)	Relative Intensity
Symmetrical Stretching (ν_s)	2870-2840 (2855)	<i>m</i>
Asymmetrical Stretching (ν_a)	2940-2915 (2925)	<i>m-s</i>
Scissoring Bend (δ_s)	1480-1440 (1460)	<i>m</i>
Twisting Bend (δ_t)	1350-1150 (1250)	<i>i</i>
Wagging Bend (δ_w)	1350-1150 (1250)	<i>i</i>
Rocking Bend (δ_r)	~720	<i>i</i>

Vibrational intensities have generally been overlooked or neglected in the analysis of vibrational spectra (Socrates 2001). 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. Although many texts list these intensities in subjective terms due to instrumentation flexibility, their relation to one another can be objective (Socrates 2001). Workman and Springsteen (1998) noted that these peaks are generally calibrated based on a linear relation to the area under the peak or a polynomial relationship to the

height of the peak. Based on these observations, a quantitative description of the size and shape of absorption bands is proposed in Table 9.2.

Table 9.2: Approximations of Transmittance of IR/NIR Active Tones (Socrates 2001; Workman and Weyer 2008).

Relative Intensity	Peak Transmittance	Transmittance Shape
<i>s</i>	10%	Normal
<i>m-s</i>	30%	Normal
<i>m</i>	50%	Normal
<i>m-w</i>	70%	Normal
<i>w</i>	90%	Normal
<i>br</i>	n/a	Inverse Beta (square)
<i>sh</i>	n/a	½-Var. Normal
<i>d</i>	n/a	2, ½-Var. Normal
<i>i</i>	100%	n/a

The peak intensities are scaled from 0-100% transmittance, depending on the level of absorbance. A *w* indicates 90% transmittance, *m-w* a 70% transmittance, and so on.

Rule 19. The peak intensities are scaled according to transmittance, as shown in Table 9.2.

The shape for the absorbance/transmittance of the tone is determined using probability theory, where normal distributions are fitted to the mean of each transmittance, with the ranges of the wavelength region listed in the texts marking the 95% probability interval. Thus, the σ standard deviation for the response is calculated as:

$$\sigma = \frac{\lambda^{UL} - \lambda^{LL}}{4} \quad (9.5)$$

where P^{RANGE} is the difference between the expected regions of the upper and lower absorbances/transmittances. The transmittance probability values are given as a normal distribution:

$$f^{abs} = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(\lambda - \lambda^{bar})^2}{2\sigma^2}} \quad (9.6)$$

where λ is the wavelength where the absorbance/transmittance occurs, λ^{bar} is the peak wavelength location, and f^{abs} is the probability density function. These values can then be scaled using the peak transmittances associated with the tone.

$$\frac{P_{peak}^T - P_{2\sigma}^T}{f_{peak}^{abs} - f_{2\sigma}^{abs}} = \frac{P_{peak}^T - P^T}{f_{peak}^{abs} - f^{abs}} \quad (9.7)$$

For situations where no range is listed, such as for the rocking-bend vibration of 720cm^{-1} in Table 9.1, a range of $\pm 10\text{cm}^{-1}$ is artificially applied.

Rule 20. A normal distribution is used to approximate the shape of an absorbed tone, spanning the range listed in the text. For situations where no range is listed, a range of $\pm 10\text{cm}^{-1}$ is artificially applied.

Occasionally, the transmittance (absorbance) is noted as *sh* sharp. For this situation the range is narrowed by reducing the equivalent standard deviation by half. When the transmittance tone is listed as a *br* broad absorbance, an inverse beta distribution, defined by two positive shape parameters, known as α and β which are each set to 0.5, is utilized. For simplicity this distribution is best approximated using a square estimation, spanning the identified range. Occasionally, the absorbance presents as a *d* doublet, which can be handled by splitting the distribution into two, half variance normal distributions centered around the doublet peaks.

Table 9.2 lists various approximations of the transmittance types noted by Socrates (2001) and Workman and Weyer (2008).

Rule 21. A normal distribution with half the variance is used to approximate the shape of sharp tone and an inverted beta distribution with $\alpha=\beta=0.5$ is used to approximate the shape of a broad tone.

Once the fundamental tones have been identified, the overtones can be estimated from Workman and Weyer (2008). As noted in Table 3.1, beyond the 2nd overtone the relative intensities of absorptions are sufficiently small as to be negligible when combined with the fundamental tones found in the IR frequency range.

Rule 22. Using Eq. 3.33, the locations of the 1st and 2nd overtones are calculated for the fundamental tones identified in Rule 18.

The overtones are then cross-referenced to find the physically active tones. For the CH₂ methylene group, Workman and Weyer (2008) identify a 1st overtone peak of asymmetrical stretching ($2\nu_a$) and symmetrical stretching ($2\nu_s$) at 5671 cm⁻¹ and 5680cm⁻¹, respectively (Tosi and Pinto 1972; Buback and Harfoush 1983). The remaining scissoring bend first overtone ($2\delta_s$) in Table 9.1 was inactive according to Socrates (2001). For the second overtone region, both stretching vibrations ($3\nu_s$, $3\nu_a$) occur between 8389cm⁻¹ to 8247cm⁻¹, depending on the influence of any methyl groups in the sample, which is addressed through higher ordered group contributions (Workman and Weyer 2008). The scissoring bend 2nd overtone ($3\delta_s$) was inactive. This result was expected since the 1st overtone band was also inactive, and, according to Table 3.1, the intensity was expected to be even smaller for the higher overtone. This observation is stated as Rule 23.

Rule 23. Once a tone has been labeled inactive, all higher ordered tones and combination tones utilizing the inactive tone shall also be considered inactive. This rule will not apply to combinations involving two different bond tones, even if one of them is inactive.

In addition, the proximity of fundamental asymmetric and symmetric stretching tones also gives rise to a strong combination band ($\nu_a + \nu_s$) at 5800cm^{-1} (Buback and Voegelé 1993). This combination tone has been shown to move slightly in linear aliphatic compounds due to the influence of methyl groups, but can be accounted for using higher order contribution corrections which are later described in Eq. 9.8 (Workman and Weyer 2008). Other combination bands include fundamental stretching-bending combinations of $\nu_s + \delta_s$ and $\nu_a + \delta_s$ at $4336\text{-}4332\text{ cm}^{-1}$ and $4257\text{-}4262\text{ cm}^{-1}$, respectively. Also, a doublet from a first overtone uncoupled stretch plus a fundamental bending combination is also prevalent at 7186cm^{-1} and 7080 cm^{-1} (Murray 1987). A number of weaker combination bands containing inactive fundamental bands can be found in the area between the first combination and first overtones bands (Workman and Weyer 2008). The final, combined IR & NIR methylene CH_2 group spectrum is shown in Table 9.3. A list of other groups can be found in Appendix A6.

Table 9.3: The Derived IR Absorption Spectrum for the CH_2 Methylene Group.

Band	Wavelength Region (cm^{-1})	Relative Intensity & Shape
Symmetrical Stretching (ν_s)	2870-2840 (2855)	<i>m</i>
Asymmetrical Stretching (ν_a)	2940-2915 (2925)	<i>m-s</i>
Scissoring Bend (δ_s)	1480-1440 (1460)	<i>m</i>
1 st Overtone Asym. Str. ($2\nu_a$)	5681-5661 (5671)	<i>w (in NIR)</i>

1 st Overtone Sym. Str. ($2\nu_s$)	5690-5670 (5680)	w (in NIR)
2 nd Overtone Asym. Str. ($3\nu_a$)	8399-8379 (8389)	m-w, sh (in NIR)
2 nd Overtone Sym. Str. ($3\nu_s$)	8257-8237 (8247)	m-w, sh (in NIR)
Combination ($\nu_a + \nu_s$)	5900-5700 (5800)	m-w (in NIR)
Combination ($\nu_s + \delta_s$)	4336-4332	s (in NIR)
Combination ($\nu_a + \delta_s$)	4262-4257	s (in NIR)
1 st Overtone Comb ($2\nu + \delta$)	7186 - 7080	m-w (in NIR)

One of the important observations concerning vibrational spectroscopy is that the strength of the absorption is dependent on both the type of vibration and the number of times the vibration appears in the molecular architecture (Socrates 2001). This conclusion lends itself easily to concept of additive absorption peaks. Unfortunately, as noted earlier, not all intensities are directly additive; some locations of absorptions move and some intensities only change at a fraction of what is expected. That is because the intensity of the band may also indicate the presence of certain atoms or groups adjacent to the functional group responsible for the absorption band (Socrates 2001). To account for this behavior, 2nd, 3rd, and other higher order group contributions can be utilized to provide corrections to the 1st order group contributions. Marrero and Gani (2001), amongst many others (Joback and Reid 1983; Joback and Reid 1987; Constantinou et al. 1996), identified several 2nd and 3rd order group contributions typically found in thermodynamic databases. As discussed in Chapter 8, a method to identify the 2nd order groups from 1st order groups was developed by Chemmangattuvalappil *et al.* (2009) and Solvason *et al.* (2009) to aid the product developer. These techniques can also be used in the development of IR/NIR characterization groups. For example 2-propanol contains the following 1st order groups: 2 methyl CH₃ groups, 1 methine CH group, and 1 hydroxyl OH group. It also

contains the 2nd order secondary alcohol CHOH group. The 2nd order group is necessary since the hydroxyl OH group induces a strong dipole in the methine CH group and, as a result, the location and intensities of the IR and NIR absorptions are affected.

Although an effective method of correcting for interactions between groups, simply developing the IR/NIR parameters for only the groups identified in the GCM limits the scope of this method to the description of molecular architecture to that of small molecules, and polymer backbones. In order to expand the method to include other architecture information provided by IR/NIR spectroscopy, such as information on inter- and intra-molecular hydrogen bonding (Socrates 2001), polymorphic crystal structure (Salari and Young 1998), and many others, it is necessary to reformulate the group contribution approach in terms of Eq. 9.7. In this formulation, the first term is the 1st order group contribution and the second term κ_i is a combination of the corrections from all higher order contributions, including those specific to molecular architecture existing at larger length scales.

$$f(X) = \sum_i N_i (C_i + \kappa_i) \quad (9.8)$$

The parameters of the corrections can be estimated in two ways. In the first method, the fundamental, combination, and overtones of each higher order as defined by Chemmangattuvalappil *et al.* (2009), Marrero and Gani (2001), and/or the researcher can be estimated using the rules outlined earlier in this section. Although this result gives accurate results, it is quite tedious. The second method is simpler, but has greater potential for error and can lead to degeneracy. In this method, progressively larger groups completely contain the information of the smaller groups, but also contain corrections for 2nd order, 3rd order, structural, and orientation effects. The specific differences between the orders of the groups are not quantified directly, but rather different groups containing various levels of information are

developed. The groups are hierarchical in nature, with the groups containing the higher ordered information preferred over the groups containing lower orders. For example, in the determination of the IR/NIR spectra for 2-propanol, a second order CHOH group would completely replace the CH and OH first order groups since it already contains their information; plus it contains additional information on the induced dipole interaction.

Rule 24. A hierarchical group structure is created such that higher order groups contain interaction information plus lower order group information.

As a result, the best model of the molecular architecture is achieved by specifying the largest functional group first, then the second largest, and so on. In most cases, a higher order functional group will completely overlap lower order functional groups.

Rule 25. When one group completely overlaps another group, the larger group will be chosen for the design. As a result, the largest functional group is specified first, the next largest second, and so on.

In some situations only partial overlaps may occur. In the GCM method proposed by Marrero and Gani (2001) and reviewed in Chapter 8, 1st order groups are not allowed to partially overlap while 2nd order and higher groups are allowed to partially overlap. The assumption made by Marrero and Gani (2001) is that any error from double counting of 2nd order effects is outweighed by the errors associated with failing to account for second order effects when they do, in fact, exist. For the characterization based group contribution method (cGCM) proposed here, no partial overlaps are allowed since the contributions from 1st and 2nd order and higher groups cannot be separately adjusted without rewriting the current state-of-the-art combinatorial algorithm. Since the method will fail to specify any corrections to the first order combination at

the overlap interface, groups are built such that the combinations occur across the minimum number of C-C bonds (Marrero and Gani 2001).

Rule 26. Candidates with partial overlaps should be reconfigured to lower order groups containing the minimum number of bond connections and maximum amount of information.

For example, isobutanol can be built from a multitude of groups, including combinations of the following: a methyl CH_3 group, methylene CH_2 group, a methine CH group, an alcohol OH group, a secondary alcohol CHOH group, a primary alcohol CH_2OH group, and a terminated primary alcohol $\text{R}_1\text{R}_2\text{CH}_2\text{OH}$. Figure 9.3 outlines 6 potential methods of combining the groups to build isobutanol. The first method uses only 1st order groups, and has the minimum amount of information available. The second and third methods use a combination of 1st and 2nd order terms, introducing some interaction corrections. The 4th method incorporates some 3rd order information, but contains two C-C bonds over which no interaction information can be obtained. The 5th method represents the best available combination, using only one C-C bond and two 2nd order groups. Finally, although the 6th method combines a 2nd order and 3rd order group and has only 1 bond connection, it also possesses a partial overlap which cannot be handled with the cGCM algorithm although it is handled in the traditional GCM algorithm (Marrero and Gani 2001).

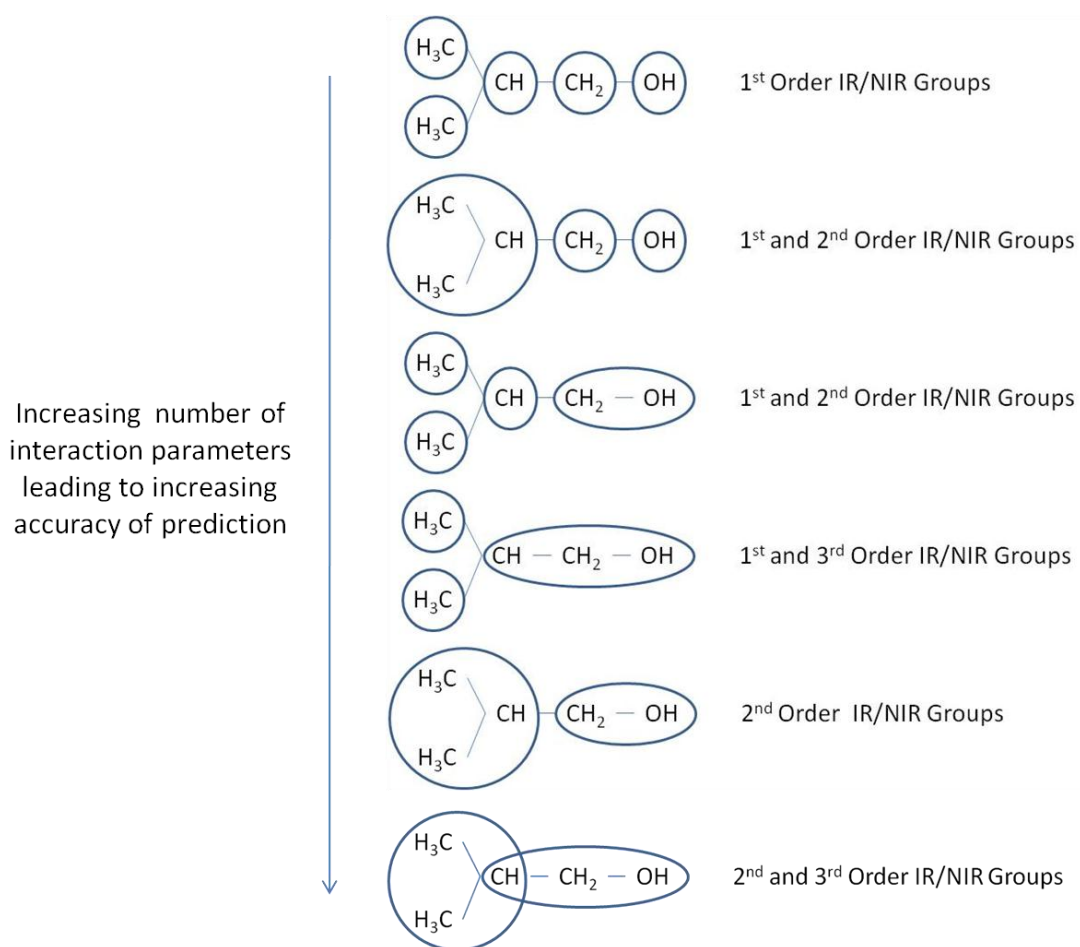


Figure 9.3: Potential Combinations of IR/NIR Groups for Building Isobutanol.

Occasionally, bonds will need to be built across other atoms, like the C-O bond in an aliphatic ether (-O-). For this reason, an additional constraint is applied when building molecules involving this group such that it must be attached to the carbon atom in other groups.

Rule 27. Candidate molecules can also be built over aliphatic ether groups (-O-), which contain bond information between the O atom and two C atoms. As such, the ether groups can only be attached to the carbon atoms of other groups.

One of the benefits of using characterization based combinatorial methods is the ability to handle information unique to the characterization. For instance, IR/NIR spectroscopy not only

provides information on the functional groups within the molecule, but also can provide information on stereoregularity (Zerbi *et al.* 1965; Socrates 2001; Solvason *et al.* 2010), hydrogen bonding (Desiraju 1996; Workman and Weyer 2008; Izutsu *et al.* 2009), solid particle shape/size (O'Neil *et al.* 1998; Ndindayino *et al.* 1999; Sorensen *et al.* 2006; Almaya and Aburub 2008), polymorphism (Deeley and Spragg 1991; Kondo 1997; Norris *et al.* 1997; Salari and Young 1998; Yu *et al.* 1998; Nikonenko *et al.* 2002; Abebe *et al.* 2008) and crystallinity (Zerbi *et al.* 1965; Salari and Young 1998; Seyer *et al.* 2000; Stephenson *et al.* 2001). This extra structural information can be made to enhance the understanding of the molecular architecture. For instance, O'Neil *et al.*(1998) demonstrated how bulk properties such as particle size and porosity can influence IR/NIR spectra. In agreement with Mie theory, the reflectance of the IR spectra varied with the mean particle size, although the particle shape also had a large influence (Ciurczak 1987; O'Neil *et al.* 1998). In particular, the baseline NIR reflectance elevation was observed to decrease as the particle diameter increased (Abebe *et al.* 2008). Abebe *et al.* (2008) also noted that the more non-spherical the particle is, the more muted the elevation of the NIR spectra. Beyond the bulk properties of the particle size and shape is its microstructure. In general, the particles may be of granule or crystalline form. In the granule form, multiple, alternating layers of amorphous and crystalline regions are present. The crystalline form is generally treated as a single polymorph. Although some progress relating specific crystallinity bands to particular NIR was made by Zerbi (1965), the results still remain case specific. At the present time, x-ray diffraction (XRD), wide angle x-ray spectroscopy (WAXS), and small angle x-ray spectroscopy (SAXS) remain the preferred methods for developing the crystalline structure.

As opposed to the group combinatorial approach outlined earlier, there is not a distinct hierarchy for which to add a diameter effect, nor is there a distinct particle size at which a particular IR effect is observed across all molecules. Rather the effect is relative to the overall baseline reflectance-absorbance on a case-by-case basis. This is beneficial to the combinatorial approach used in mixture and molecular design. For example, diffuse reflectance is one of the six ways in which an incident EM wave can interact with a material; it can also be absorbed, transmitted, diffuse transmitted (i.e. scattered), specularly reflected (i.e. regular reflected), and retroreflected back to the source (Workman and Springsteen 1998). In most instances for IR and NIR spectroscopy, only three of the modes are observed, transmittance, absorbance, and diffuse reflectance. In most cases, a reflected-absorbance response (e.g. Fig. 9.4) is obtained by applying Eq. 3.29 to the reflectance (P^R) instead of transmittance (P^T). It has been shown that NIR wavelength diffuse reflectance exhibits an inverse relationship with mean particle size in agreement with Mie theory (O'Neil *et al.* 1998). The reason for this is that as the particle diameter increases, the total volume of the sample does not, meaning that void space increases in the sample which, in turn, decreases the pathlength of the EM signal and its resulting reflected-absorbance. In similar fashion, the increase in void space also results in fewer solid particles reflecting the EM signal, leading to a lower diffuse reflectance. This phenomenon is sometimes referred to as baseline elevation, which results from systematic error inherent in the spectrophotometer, and is handled using linear regression baseline fitting. For particle size d_p , O'Neil *et al.* (1998) and Abebe *et al.* (2008) showed that a line can be fit to two regions of the spectrum in order to calibrate the spectral response:

$$\ln d_p = b_o + b_1 P_1^R + b_2 P_2^R \quad (9.9)$$

where each regression coefficient is species dependent and the two centered reflectance wavelengths (P^R) were chosen by searching for the combination that provided the minimum error of calibration for a Kubelka-Munk function (O'Neil *et al.* 1998). The quadratic least squares inverse of Eq 9.9 is most useful and can provide particle diameter corrections for specific types of molecular architecture. In particular, models can be built for complex mixtures when only some of the constituent concentrations are known; the only requirement being the selection of the appropriate wavelengths that correspond to the reflected-absorbances of the desired molecular architecture (Workman and Springsteen 1998). This requirement results because the baseline elevation of the intensity of the reflected-absorbance response is not constant across all wavelengths as shown in Fig. 9.4 (Abebe *et al.* 2008).

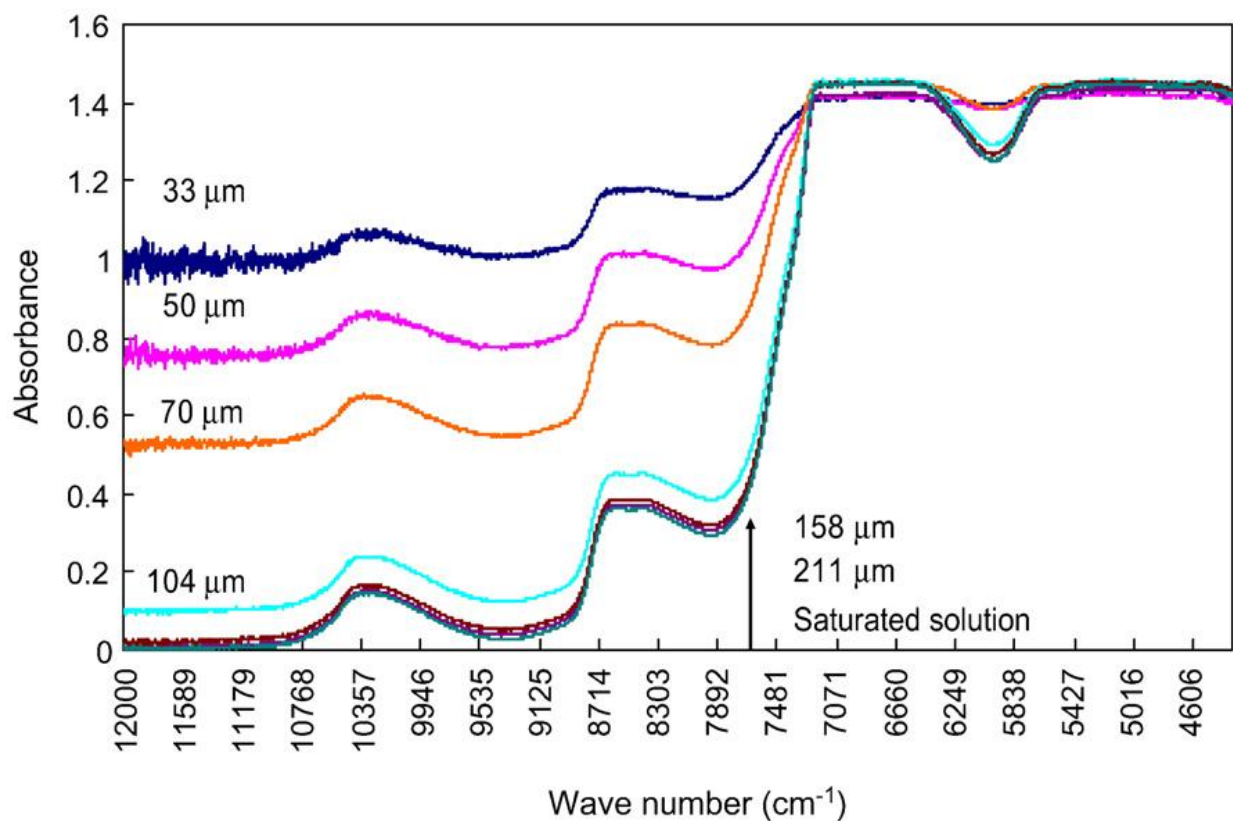


Fig. 9.4: Raw, Unscaled NIR Reflected-Absorbances of Slurries of L-Glutamic Acid Crystals at Different Diameters (Abebe *et al.* 2008)

As a result, corrections for particle diameter may be assigned to specific group reflected-absorbances provided baseline elevation expressions or responses are available. Unfortunately, in order to ensure a unique description of the inverse of Eq. 9.9, the number of samples used in the training set, usually 30-60, would then represent the maximum number of descriptors that could be fit. In otherwords, a full description of the spectroscopic molecular architecture would require a prohibitively large training set which, by extension, would introduce colinearity into the solution resulting in an over-fit and poor predictive description. A better way to capture this information is to make use of eigenvector quantitation (i.e. decomposition) to find a latent sub-domain that contains all molecular architecture information. This approach works well within the reverse problem formulation since the objective is to find a computationally efficient method of systematically determining the molecular architectures that deliver a set of desired property attributes.

9.1.2. Developing the Latent Property Domain

Although the IR/NIR characterization based group contribution can handle some interaction effects within the group structure, the colinearity of larger data sets require the use of domain reduction techniques. In addition, the IR/NIR characterization based group contribution method of defining groups in terms of a hierarchical structures means that enumeration will result in multiple candidates with identical molecular architectures with varying levels of information. Although it has been shown that this degeneracy can be handled using a set of rules that help select the best representation of the candidate molecule, separating the effects into independent spectra is difficult. Furthermore, as the size of the molecule grows, increasing amounts of structure based information, from simple hydrogen bonding, to crystal unit cell structure and morphology, will require an exponential, and possibly, case specific rule structure.

In order to better handle these types of interactions, it is suggested that (1) deconvolution algorithms be applied to better ascertain the effects on the IR spectra or (2) the combinatorial structure be mapped into a linear sub-domain that facilitates an easier computation. This sub-domain provides both an avenue for combinatorial efficiency and suitable variable structure from which attribute-property models can be built.

The applicability of vibrational spectroscopy techniques are constrained, to a great extent, by the low resolution of bands in the spectra of complex substances (Nadler *et al.* 1989). The spectra available for identification, such as for the methylene CH₂ group, are considerably smaller than the actual number of individual components (Nikonenko *et al.* 2002). Deconvolution algorithms serve to enhance the absorption spectra in order to better differentiate the individual absorption bands. The technique works by removing the bandpass function (similar to a Fourier transform) that naturally smoothes a spectra (Nadler *et al.* 1989). Two methods are typically applied, signal space iterative deconvolution and frequency space deconvolution (e.g. Fourier transform). Although a strong technique, it struggles from subjective interpretation of spectra. In contrast, using a decomposition algorithm to develop a sub-domain uses the colinear and convoluted nature of the absorption spectra to build a strong latent property structure. The choice of the appropriate training set spectra is vital to building the latent variable structure since the training set defines the molecular architecture building blocks that can be used in the characterization based GCM (cGCM). The training set is also used to handle unusual interactions that cannot be described by characterization based cGCM. These interactions may include non-equilibrium thermodynamic states, instrumentation bias like temperature drift, and many other types of systemic error. Decomposition algorithms like PCA help to neutralize these types of interactions by using the variance-covariance structure to compress the most important

molecular architecture data to an underlying latent variable structure while limiting the impact of the unusual interaction. As a result, only the interactions identified by the cGCM structure as estimated from the data are included in the design in the latent subdomain.

Rule 28. Only the interactions identified by the cGCM structure as estimated from the training set data should be included in the design.

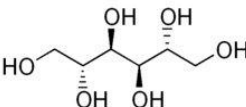
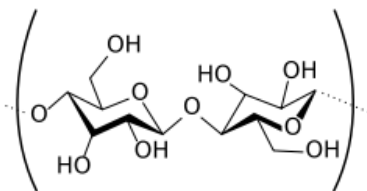
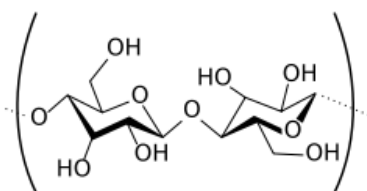
For this method to be successful, multivariate normality of the data structure must be verified. To test the non point estimate of multivariate normality, the square of the generalized mahalanobis distance is evaluated using the chi square distribution (Johnson and Wichern 2007). Under sufficient conditions for normal populations, point estimates of the system can replace the non point estimates. The distribution that best approximates this condition is Hotelling's T^2 distribution. Multivariate normality is checked using a gamma Q-Q plot of the residuals against the quantiles of the system. When the line is straight, has a slope of 1 and, contains few outliers, then it is said to be multivariate normal. If the system is found to not be multivariate normal, various linearization functions must be applied, such as mean square, logarithmic, Fisher correlation, and power transformations, among others (Johnson and Wichern 2007). Once the system is validated as multivariate normal, it must be centered and standardized in order for the decomposition to be at its most effective. This is accomplished by subtracting the mean of each property and dividing by the square root of the variance of each property (Gabrielsson *et al.* 2002; Johnson and Wichern 2007).

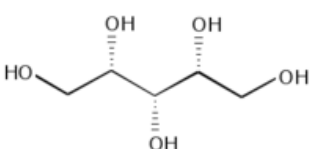
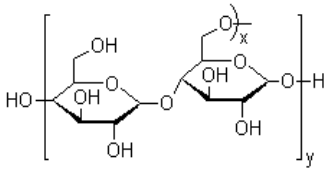
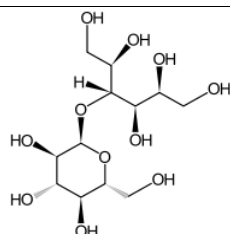
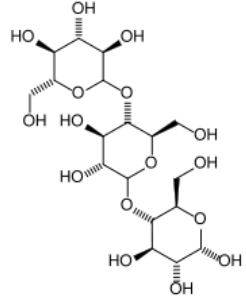
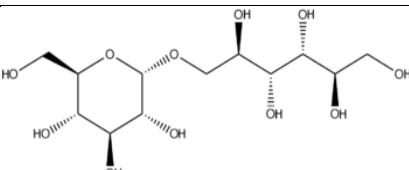
Rule 29. The molecular architecture characterization data must be verified to be multivariate normal before any statistical techniques are applied.

Rule 30. The molecular architecture characterization data should be centered and standardized prior to analysis with decomposition.

Sometimes, when two different characterization data sources are combined, it is better to standardize and center them separately. This helps to minimize the systematic noise in the system that results from the choice of characterization equipment. Examples include IR and NIR spectroscopy (Gabrielsson *et al.* 2002; Gabrielsson *et al.* 2003; Gabrielsson *et al.* 2004; Gabrielsson *et al.* 2006). Dissimilar characterization methods may require a more rigorous analysis. For example, in the design of an excipient filler for a direct compressed acetaminophen tablet, Gabrielsson *et al.* (2003) evaluated 28 potential candidates using both IR and NIR spectroscopy. The candidate fillers can be found in Table 9.4 (Gabrielsson *et al.* 2003).

Table 9.4: Filler Candidates for a Pharmaceutical Excipient in a Direct Compressed Acetaminophen Tablet (Gabrielsson *et al.* 2003)

Fillers <i>i</i>	Common Name	Type of Filler	Molecular Structure	Particle Diameter (μm) and Structure*
1	Pearlitol SD200	Mannitol		200, g
2	Pearlitol 300DC			250, g
3	Pearlitol 400 DC			360, g
4	Pearlitol 500 DC			520, g
5	C Mannidex			N/A, g
6	Avicel PH- 101	Microcrystalline Cellulose		50, c
7	Avicel PH- 102			90, c
8	Avicel PH- 112			90, c
9	Avicel PH- 200	High Density Microcrystalline Cellulose		180, c
10	Avicel PH- 301			50, c
11	Avicel PH- 302			90, c

12	Xylitab DC	Xylitol		N/A, g
13	Xylisorb 90			90, c
14	Xylisorb 300			250, c
15	Xylisorb 700			700, c
16	Xylitab 100			200, g
17	Xylitab 200			300, g
18	Xylitab 300			150, g
19	C Xylidex CR 16055	Maltodextrin		200, g
20	Lycatab			300, g
21	C Sperse MD 01314			150, g
22	C Pharm 01980			N/A
23	C Pharm 01983			N/A
24	Maltisorb P90	Maltitol		N/A
25	Maltisorb P200			N/A
26	Promogram W	Dextrin		N/A
27	Isomalt DC 100	Isomalt		N/A
28	Isomalt DC 110			90, N/A

* Structure types are *g* granular or *c* crystalline.

The IR and NIR spectroscopic images of the first five fillers, derivatives of Mannitol, are provided in Figure 9.5. Transmittance IR spectroscopic data was measured between 4500 to 650 cm^{-1} and represented as 1997 separate wavelengths, approximate 2 cm^{-1} apart. Reflected-absorbance NIR spectroscopic data was measured between 400 and 2500 nm and represented 1050 separate wavelengths, approximately 2 nm apart. A table of the IR and NIR spectroscopic responses for all filler excipients in Table 9.4 is available in Appendix 5.

The IR and NIR data sets were then standardized and centered separately using Eq. 9.10 and 9.11 and then combined to form the characteristic spectroscopic profile shown in Fig. 9.6.

$$\hat{P}_{ji}^{abs} = \frac{P_{ji}^{abs} - P_{avg,i}^{abs}}{\sqrt{S_i^{2,abs}}} \quad (9.10)$$

$$\hat{P}_{ji}^T = \frac{P_{ji}^T - P_{avg,i}^T}{\sqrt{S_i^{2,T}}} \quad (9.11)$$

It should be noted that number of $j\epsilon p$ absorbance wavelengths are much larger than the number of $i\epsilon u$ samples and appear highly correlated, making them strong candidates for decomposition. Testing for multivariate normality in Fig. 9.7 reveals that the system behaves in a multivariate normal manner, except for the last two filler excipients of maltodextrin, runs 27 and 28, which appear to be outliers in the dataset. These values may indicate a nonlinear relationship in the data that may artificially influence the definition of latent variable structure. In spite of this observation, investigation of the IR and NIR spectra yielded no insight into the cause. Further analysis of the spectroscopic properties P is warranted, but in the interest of matching the results published by Gabrielsson *et al.* (2003) these data points were included.

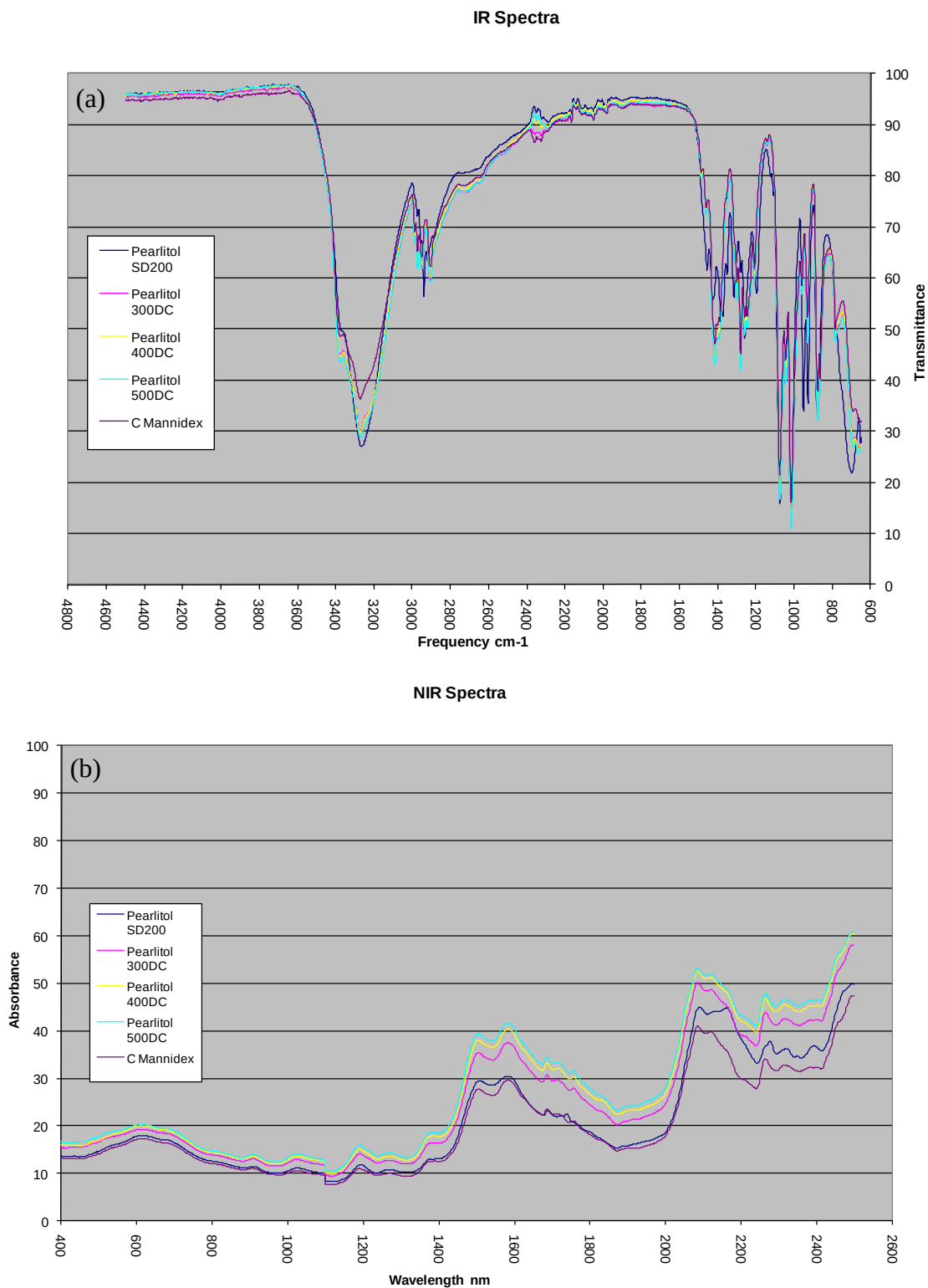


Figure 9.5: (a) IR Transmittance and (b) NIR Reflected-Absorbance Spectra for Mannitol (Gabrielsson *et al.* 2003).

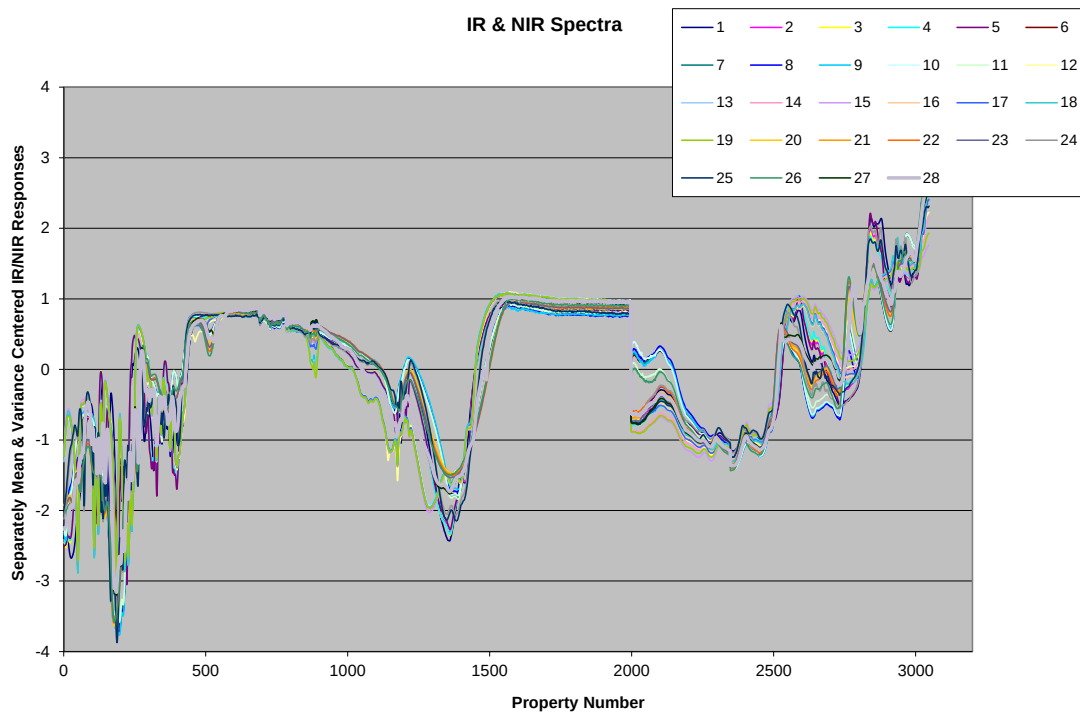


Figure 9.6: p -Standardized and Centered IR Transmittance and NIR Reflected-Absorbance filler excipient spectra (Gabrielsson *et al.* 2003).

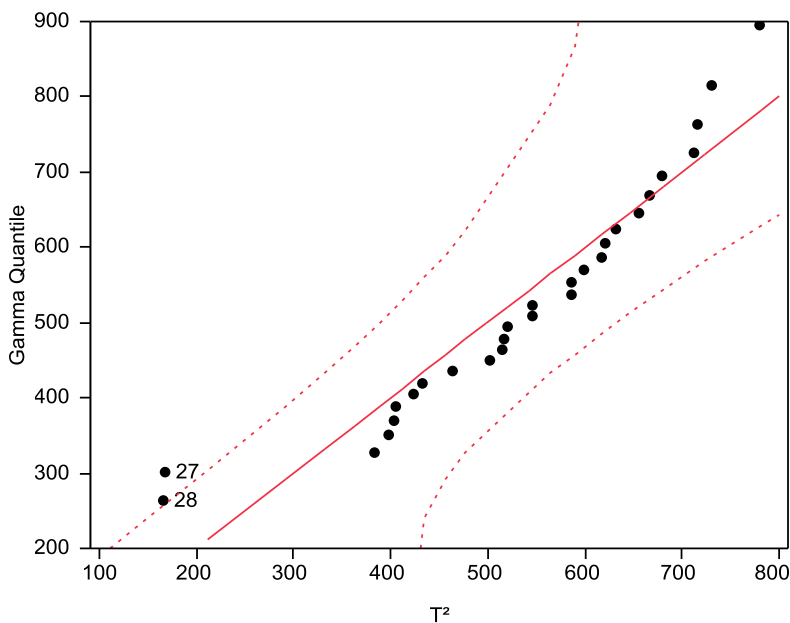


Figure 9.7: The Gamma Q-Q Plot for the Filler Excipients Presented by Gabrielsson *et al.* (2003)

Once the system is determined to be multivariate normal, a variety of decomposition techniques exist to aid in the reduction of the number of jep properties describing the system to a

set of tem latent variable scores and lem latent variable loadings. These techniques include singular value decomposition (SVD), spectral decomposition, and the NIPALS algorithm and can be found in a variety of commercial softwares, such as JMP, SAS, MatLab, and SIMCA-P. Decomposition by principal component analysis (PCA) is the most common method of finding the latent property subdomain. By definition, PCA uses the variance-covariance structure to compress the molecular architecture data to principal component data that contains much of the system variability. This result can improve the interpretation of the data structure by consolidating multiple molecular architecture effects into single, underlying latent variables which are devoid of colinearity. It is generally considered acceptable to apply decomposition techniques to spectroscopy based characterizations because the main molecular architecture information is collected in the larger population l loadings while the influence of that information is located in the scores t (Naes and Isaksson 1992). If the original property data of molecular architecture descriptors are of the same type, then the intensities of the absorbances can be decomposed into loadings L_{pxm} , which represent the underlying latent structure of the data (Johnson and Wichern 2007) and scores T_{nxm} , which represents the weightings of each latent spectra (Workman and Springsteen 1998).

$$P_{n \times p} = T_{n \times m} L'_{m \times p} \quad (3.35)$$

As noted in Chapter 3, the first 2 or 3 principal components typically account for 80% to 90% total variance. The remaining components can be removed without much loss of information (Johnson and Wichern, 2007). While this approach represents a significant gain in efficiency, the drawback is that this reformulation is not unique. For instance, a decomposition of the IR/NIR spectra of the fillers in Figure 9.6 by the SIMCA-P program is given in Table 9.5. The program uses the NIPALS algorithm and the analysis was performed on both the covariance

and correlation matrices. Three principal components were found to contain 87% of the data and are presented in Table 9.5.

Table 9.5: The Principal Components for the Filler Excipients using SIMCA-P Software on the Covariances and Correlations (Gabrielsson *et al.* 2003).

Sample	T_1^{Cov} (SIMCA-P)	T_2^{Cov} (SIMCA-P)	T_3^{Cov} (SIMCA-P)	T_1^{Corr} (SIMCA-P)	T_2^{Corr} (SIMCA-P)	T_3^{Corr} (SIMCA-P)
1	-4.4	-42.8	17.3	-4.0	-42.4	-22.2
2	7.2	-35.9	15.5	7.5	-35.8	-19.8
3	7.5	-36.8	17.0	7.8	-36.6	-21.4
4	10.5	-35.6	15.0	10.7	-35.6	-19.3
5	1.2	-38.8	15.1	1.6	-38.4	-19.3
6	-53.6	6.1	-21.6	-52.8	3.2	21.6
7	-55.3	5.0	-24.3	-54.5	1.8	24.1
8	-57.9	-2.6	-33.1	-57.1	-6.5	32.2
9	-57.2	4.9	-21.9	-56.3	2.0	21.9
10	-48.0	7.0	-26.0	-47.3	3.6	26.0
11	-42.3	5.1	-20.2	-41.8	1.8	19.8
12	53.6	19.5	-9.5	53.0	17.3	10.7
13	53.6	9.9	-19.3	53.1	6.9	19.6
14	55.3	5.4	-14.6	54.7	2.6	14.1
15	61.8	10.9	-8.4	61.2	8.6	8.4
16	48.5	9.6	-16.4	48.1	6.9	16.6
17	49.8	11.8	-13.2	49.4	9.5	13.7
18	47.6	5.0	-23.4	47.2	1.7	23.0
19	61.8	11.9	-8.4	61.2	9.7	8.7

20	-26.8	20.1	19.2	-26.5	20.3	-17.9
21	-23.4	25.3	22.4	-23.2	25.8	-20.6
22	-27.9	30.1	26.4	-27.6	31.1	-23.9
23	-22.0	24.3	28.3	-21.8	25.3	-26.5
24	-16.0	-34.4	-11.2	-15.6	-37.4	6.4
25	-10.9	-32.8	-5.8	-10.6	-35.6	0.8
26	-28.1	37.5	15.8	-27.8	37.8	-12.3
27	7.1	21.1	32.4	7.1	22.8	-30.5
28	4.3	27.3	36.7	4.2	29.5	-33.9

The corresponding eigenvector loadings L_{pxm} for the scores T_{nxm} , shown in Table 9.5 are given in Appendix A5. By convention, the eigenvector loadings capture spectral variations that are common across the training set and the scores describe variation within the training set and can be used to predict the molecular architecture of new molecules. However, when an alternative, spectral decomposition is applied using Matlab and JMP softwares, a different set of scores are found as shown in Table 9.6. One of the major differences between the decompositions used in Table 9.5 and 9.6 is that the default analysis in SIMCA-P program finds the underlying latent structure based on the covariance structure, while the default analysis in MatLab and JMP uses a correlation structure. Since the two spectroscopic data sets were standardized and centered separately, it is believed a better analysis would be achieved by performing decomposition on the correlation data set using a normalization across the training set molecules, as shown in Eq. 9.12:

$$P_{ji}^{CORR} = \frac{\hat{P}_{ji} - \hat{P}_{j,avg}}{\sqrt{S_j^2}} \quad (9.12)$$

Table 9.6: The Principal Components for the Fillers Excipients using Matlab and JMP Softwares (Gabrielsson *et al.* 2003).

Sample	T ₁ (MatLab)	T ₂ (MatLab)	T ₃ (MatLab)	T ₁ (JMP)	T ₂ (JMP)	T ₃ (JMP)
1	-4.0	42.4	22.2	-4.0	-42.4	22.2
2	7.5	35.8	19.8	7.5	-35.8	19.8
3	7.8	36.6	21.4	7.8	-36.6	21.4
4	10.7	35.6	19.3	10.7	-35.6	19.3
5	1.6	38.4	19.3	1.6	-38.4	19.3
6	-52.8	-3.2	-21.6	-52.8	3.2	-21.6
7	-54.5	-1.8	-24.1	-54.5	1.8	-24.1
8	-57.1	6.5	-32.2	-57.1	-6.5	-32.2
9	-56.3	-2.0	-21.9	-56.3	2.0	-21.9
10	-47.3	-3.6	-26.0	-47.3	3.6	-26.0
11	-41.8	-1.8	-19.8	-41.8	1.8	-19.8
12	53.0	-17.3	-10.7	53.0	17.3	-10.7
13	53.1	-6.9	-19.6	53.1	6.9	-19.6
14	54.7	-2.6	-14.1	54.7	2.6	-14.1
15	61.2	-8.6	-8.4	61.2	8.6	-8.4
16	48.1	-6.9	-16.6	48.1	6.9	-16.6
17	49.4	-9.5	-13.7	49.4	9.5	-13.7
18	47.2	-1.7	-23.0	47.2	1.7	-23.0
19	61.2	-9.7	-8.7	61.2	9.7	-8.7
20	-26.5	-20.3	17.9	-26.5	20.3	17.9
21	-23.2	-25.8	20.6	-23.2	25.8	20.6
22	-27.6	-31.1	23.9	-27.6	31.1	23.9

23	-21.8	-25.3	26.5	-21.8	25.3	26.5
24	-15.6	37.4	-6.4	-15.6	-37.4	-6.4
25	-10.6	35.6	-0.8	-10.6	-35.6	-0.8
26	-27.8	-37.8	12.3	-27.8	37.8	12.3
27	7.1	-22.8	30.5	7.1	22.8	30.5
28	4.2	-29.5	33.9	4.2	29.5	33.9

This operation results in the correlated standardized and centered data structure shown in Fig. 9.8.

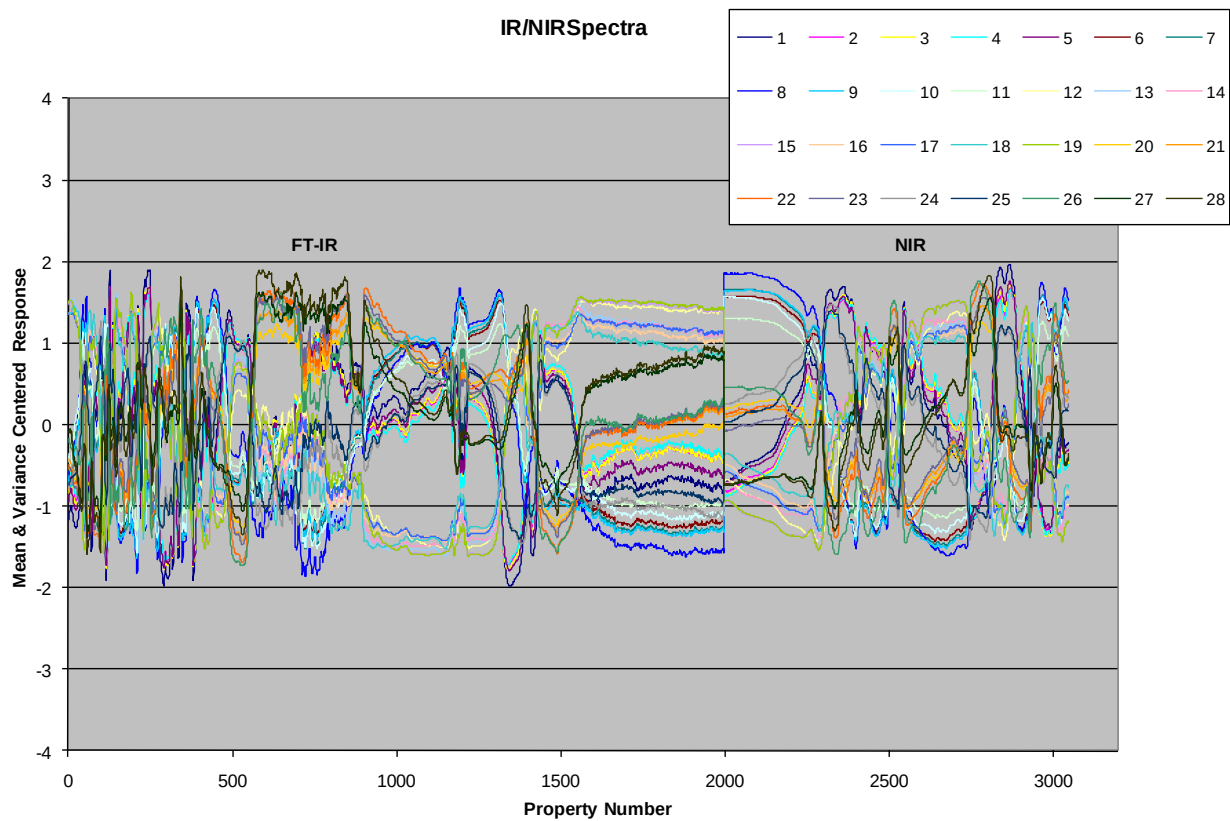


Figure 9.8: n,p -Standardized and Centered IR Transmittance and NIR reflectance-absorbance Filler Excipient Spectra (Gabrielsson *et al.* 2003).

While this transformation explains the magnitude differences between the scores in Tables 9.5 and 9.6, it does not account for the sign orientation. The sign variation between the

calculated scores of the correlations is a result of the algorithms used in MatLab, JMP, and SIMCA-P. Each algorithm arbitrarily assigns direction to either the loadings or the scores at the time of calculation. Investigations of each pair of loadings and scores in Appendix A5 reveals that the sign convention is equal and opposite for each m latent variable across all p property descriptors. Thus, when the loadings and scores are recombined, they return identical values, regardless of whether the loadings or the scores absorb the negative values. For example, the Simca-P, MatLab, and JMP predicted IR and NIR responses using only the first three correlated latent variable scores (Table 9.5 and Table 9.6) and loadings (Appendix A5) are matched against the actual response plot of Pearlitol SD200 (i.e. Mannitol) in Fig. 9.9. The responses using each of the three algorithms are identical even though the latent variable structures are not. The slight deviation from the actual response is the result of only using three of the 3047 latent descriptors to build the model, albeit the ones describing 87% of the response. In conclusion, although the non-uniqueness of decomposition prevents direct interpretation of the underlying latent variable structure, it does not prevent the interpretation of contrasts and, so long as a single set of loadings is used to map all descriptors down to the latent variable subspace, the solutions generated with the technique will be unique.

Rule 31. As long as a single set of loadings are chosen to map the problem into the subdomain, then the molecular architecture design will result in a unique solution for that given set of loadings.

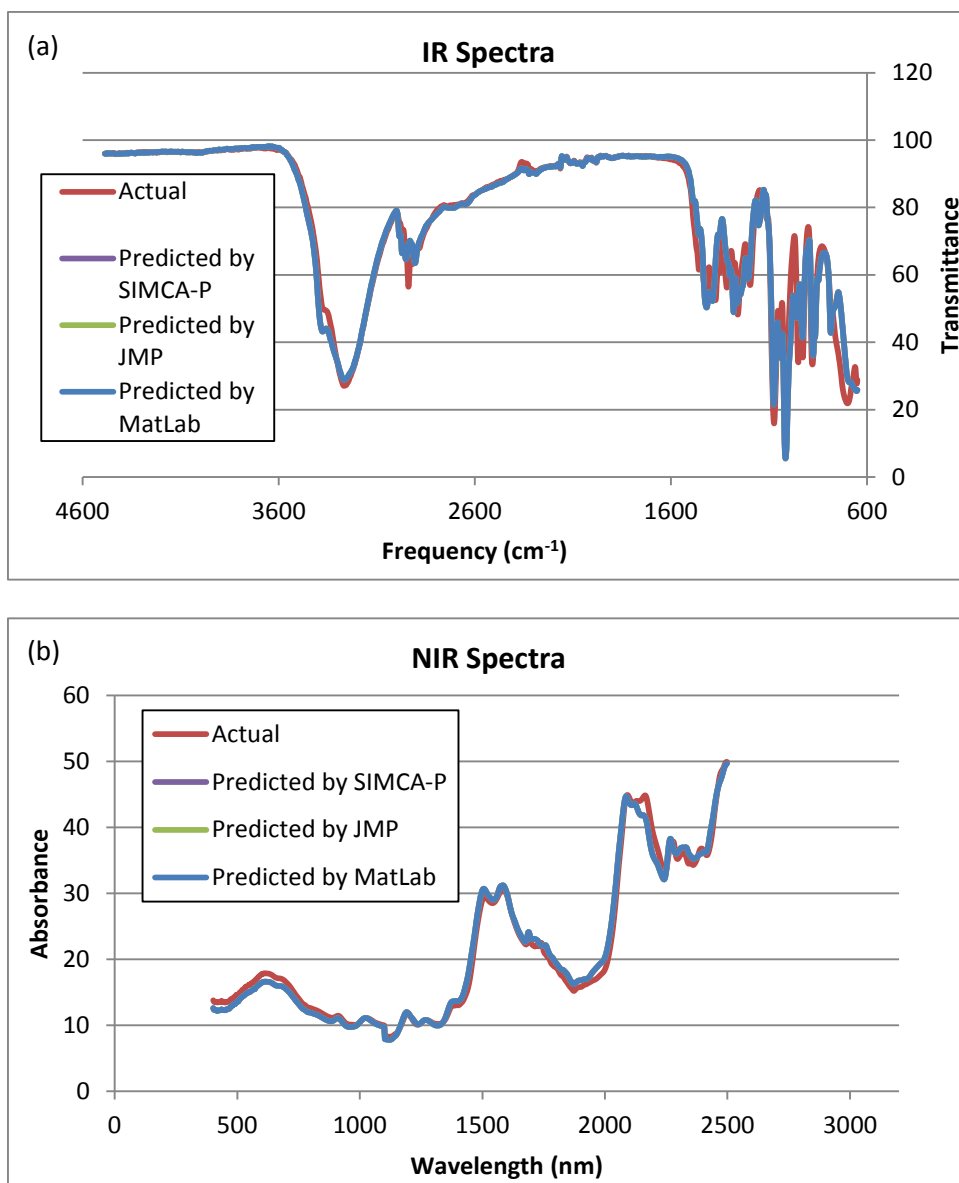


Figure 9.9: The (a) IR and (b) NIR Response plots of the Recombined Scores and Loadings Found using Simca-P, MatLab, and JMP.

In some cases, the decomposition may be artificially influenced by poorly standardized data or nonlinear influences. For example, Fig. 9.7 showed that Maltodextrin samples 27 and 28 may be overly influential in the development of the underlying architecture. This observation could have influenced the decision of Gabrielsson *et al.* (2003) to base the latent variable structure on the covariance structure and not the correlation structure of the spectroscopic

property in order to minimize their impacts. However, the convention is to base a decomposition analysis on correlation data structure when the scales of the characterization data are considered dissimilar, which they were in this instance. Therefore, in this author's opinion, this analysis should have been based on the correlation data structure. All decompositions in the case study reflect this opinion.

Lastly, once the latent domain is described, the latent property-attribute relationship is constructed. This relationship serves the primary function of mapping the target domain into the latent variable domain so that a combinatorial design using cGCM can be performed. The relationship between the principal component scores $T_{n \times m}$ and the attribute properties $A_{n \times a}$ is developed using a multivariate linear regress (MLR) on a new design of experiment (DOE) factorial design where the scores are varied between their high (+1) and low (-1) levels (Gabrielsson *et al.* 2002; Gabrielsson *et al.* 2003; Gabrielsson *et al.* 2004; Gabrielsson *et al.* 2006):

$$A_{n \times a} = T_{n \times m} B_{m \times a} \quad (9.13)$$

where $B_{m \times a}$ are the regressed coefficients found using MLR. The combined relationship is typically referred to as Principal Component Regression (PCR) since the independent variable is located in the latent sub-domain. As discussed in Chapter 3, one of the issues with PCR is that while the decomposition is designed to ensure a strong relationship between the absorbance properties, P , and the latent domain, it does not guarantee a strong relationship between the attributes A and the latent domain. The alternative is to integrate the two approaches, simultaneously calculating the structure of the latent domain and its relationship to the attributes of interest in an approach known as or a partial linear regression on to latent surfaces (PLS). This approach weighs spectra containing attribute-specific molecular architecture more heavily

in the latent variable domain, essentially compressing attribute information into the first few loading vectors (Geladi and Kowalski 1986; Workman and Springsteen 1998). Although an argument can be made to use a PLS model to develop the relationship between the property descriptors and the attribute domain if a heavy correlation between attribute and absorbance property descriptors exists, the technique was not chosen since it sacrifices some latent property descriptor accuracy for improved attribute-latent property descriptions. This choice is discussed in more detail in the case study in Sec. 9.4.

Regardless of the approach, the choice of model should be made by directly comparing the model fitness and predictive power; choosing the one with better measures. Model fitness can be measured with a variety of techniques. One of the most common is the R^2 fitness measure given in Eq. 3.25 which describes the percentage of variance explained by the model. In similar fashion, the predictive power of the model Q^2 given in Eq. 3.26 describes the percentage of predictive variance explained by the model.

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} \quad (3.25)$$

$$Q^2 = \frac{SS_{tot} - SS_{PRE}}{SS_{tot}} \quad (3.26)$$

The overall fitness and predicted power of each approach can be measured using Eq. 9.14 and Eq. 9.15:

$$R_{TOT}^2 = \prod_1^D (R^2) \quad (9.14)$$

$$Q_D^2 = \prod_1^D (Q^2) \quad (9.15)$$

where D is the total number of models used in the mapping function. Care must be taken to ensure that appropriate model fitness, as measured by R^2 , is not sacrificed to improve the prediction power, Q^2 . This can be achieved not only by selecting the appropriate latent domain structure, but also by selecting the appropriate molecular architecture relationship. The choice of molecular architecture relationship is partially dictated by the choice of characterization and partially by the training set. Choosing the characterization that maximizes the prediction power for a given number of experiments may be a procedure that can be written as a MINLP, but is beyond the scope of this dissertation.

Rule 32. The PLS or MLR model with the highest model fit and predictive power should be chosen to model the attribute-latent variable relationship.

Once the attribute-latent property, property descriptor-latent property, and cGCM expressions are developed, the method is ready for use in a computer aided molecular design algorithm. Since significant efficiency improvements have been achieved by combining aCAM^bD and aCAMD with property clusters and the reverse problem formulation, the cCAMD method will also utilize this solution structure.

9.2. Latent Property Clustering Algorithms

As noted in Chapter 6, property clustering is a tool used to improve the interpretation of the subspace properties by deconstructing the design problem into a Euclidean vector in the cluster domain and a scalar called the Augmented Property Index *AUP*. The clusters themselves are conserved surrogate properties described by property operators, which have linear mixing rules, even if the operators themselves are nonlinear (Shelley and El-Halwagi 2000; Eden *et al.* 2003). Methods for the application of group contribution methods for molecular design with

experimental data using property clustering have previously been developed in Chapter 8 and a new approach using characterization data was discussed in Section 9.1.1. To extend this approach to utilize the efficiency gains of decomposition by performing molecular design on the latent variables using the property clustering algorithm, it is important to recognize that the decomposition data structure follows a linear mixing rule. For example, rearranging the decomposition shown in Eq. 3.35 results in Eq. 9.16:

$$\mathbf{T}_{n \times m} = \mathbf{P}_{n \times p} \mathbf{L}_{p \times m} \quad (9.16)$$

This data structure is similar to the matrix form of the mixing expression shown in Eq. 6.1.

$$\boldsymbol{\psi}_{n \times p}^M = \mathbf{X}_{n \times u} \boldsymbol{\psi}_{u \times p} \quad (9.17)$$

If the loadings $\mathbf{L}_{p \times m}$ are thought of as the pure values of the principal components, then Eq. 9.16 suggests that the scores $\mathbf{T}_{n \times m}$ have behave in a similar manner to the predicted mixture properties $\boldsymbol{\psi}_{n \times p}^M$. In this situation the $j \in p$ properties for each $i \in u$ pure component, normally denoted as $\boldsymbol{\psi}_{u \times p}$ in a conventional mixture design, have been replaced by a set of $q \in m$ latent properties for each $j \in p$ property descriptor denoted as $\mathbf{L}_{p \times m}$. Likewise, the $e \in n$ mixtures of $i \in u$ pure component fractions, usually denoted as $\mathbf{X}_{n \times u}$ in a mixture design, are actually the $j \in p$ property descriptors responses (e.g. transmittances/absorbances for IR/NIR, etc.) for each $e \in n$ experimental design mixture, or $\mathbf{P}_{n \times p}$. Thus, the decomposition expression in Eq. 9.16 can be treated as a linear mixture of the latent variable loadings. There is, however, a concerning difference between the two methods: the mixture fractions $\mathbf{X}_{n \times u}$ sum to one across the $i \in u$ components and the multivariate data $\mathbf{P}_{n \times p}$ do not do the same across the $j \in p$ property descriptors. In order for latent variable models to be utilized in property clustering, it is

necessary to ensure this condition is met by standardizing the latent variable structure by dividing $\mathbf{P}_{n \times p}$ and $\mathbf{T}_{n \times m}$ by the sum total of the property descriptors for each experimental run, $\mathbf{S}_{n \times n}$.

$$S_e = \sum_j^p P_{je} , e \in n \quad (9.18)$$

$$\mathbf{Q}_{n \times m} = (\mathbf{S}_{n \times n})^{-1} \mathbf{T}_{n \times m} \quad (9.19)$$

$$\mathbf{R}_{n \times p} = (\mathbf{S}_{n \times n})^{-1} \mathbf{P}_{n \times p} \quad (9.20)$$

The resulting expression for the decomposition linear mixture model becomes:

$$\mathbf{Q}_{n \times m} = \mathbf{R}_{n \times p} \mathbf{L}_{p \times m} \quad (9.21)$$

where the new $\mathbf{Q}_{n \times m}$ matrix now represents standardized scores or mixtures, the loadings matrix $\mathbf{L}_{p \times m}$ remains unchanged, and the $\mathbf{R}_{n \times p}$ matrix now represents fractions of loadings whose cumulative sum is one for each run. Unfortunately, although the components sum to one, they are sometimes negative due to the mean-centering of the multivariate property data prior to PCA. The constraint that the fractions must be between 0 and 1 is removed with no effect on the associated mathematics, only on their interpretation, so long as a positive *AUP* is maintained by adjusting the reference property values. At this point, the loadings $\mathbf{L}_{p \times m}$ are the underlying latent variable domain subspace. In terms of property clusters, the relationship becomes:

$$Q_{eq} = \sum_{j=1}^p R_{ej} L_{jq} , e \in n , j \in p , q \in m \quad (9.22)$$

The normalized property operator of the latent variable properties may be obtained by dividing L_{jq} with a reference value $\psi_{ref,q}$.

$$\Omega_{jq}^L = \frac{L_{jq}}{\psi_{ref,q}} \quad (9.23)$$

The latent domain Augmented Property Index AUP^L , for each latent fragment is defined as the summation of all the u dimensionless property operators:

$$AUP_j^L = \sum_{q=1}^m \Omega_{jq}^L \quad (9.24)$$

Finally, the latent property cluster C_{ju}^L is defined as

$$C_{jq}^L = \frac{\Omega_{jq}^L}{AUP_j^L} \quad (9.25)$$

Shown in Fig. 9.10 is are the loadings found in Appendix A5 for the decomposed IR/NIR absorbances of the 28 filler excipients shown in Fig. 9.8 (Gabrielsson *et al.* 2003). For this system of $p=3047$ property descriptors and $n=28$ training set samples, 3 latent properties were found to contain 87% of the data. When plotted in two dimensions, the ellipsoid becomes a series of ellipses that are indicative of a well explored, multivariate normal domain. As expected, the plot of the two latent properties containing the most property descriptor information, L_1 vs. L_2 in Fig. 9.10(a), has a well-defined elliptical shape while Fig. 9.10(c), which contains latent properties with less descriptor information, is less defined.

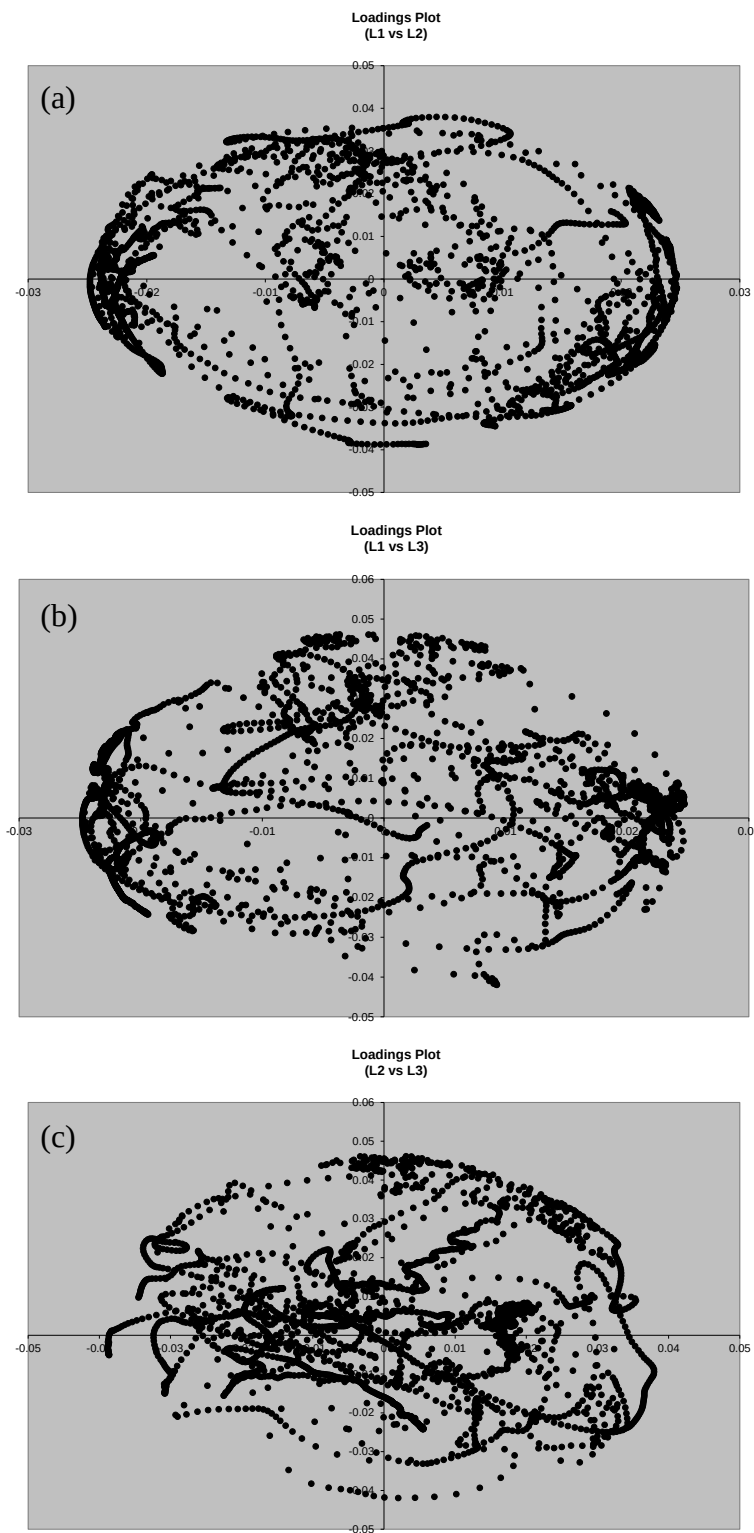


Fig. 9.10: From the PCA decomposition of the Training Set Excipient Property Descriptors shown in Fig. 9.8, Plots of the Loadings: (a) L_1 vs. L_2 , (b) L_1 vs. L_3 , and (c) L_2 vs. L_3 .

Using property clustering, these three figures are consolidated into a single chart shown in Fig. 9.11, where the x and y axis are the Cartesian coordinates of the cluster domain. The black circles typically used to denote pure components have been changed to pluses and minuses to illustrate two sets of loadings: those with positive *AUP* and those with negative *AUP*. The negative *AUP* can be made positive by changing the reference values, but for ease of transparency, the reference values were all left at unity. The figure also shows a good distribution of the loadings in each of the 8 cluster regions identified in Chapter 7 which indicate a strong description of the latent domain.

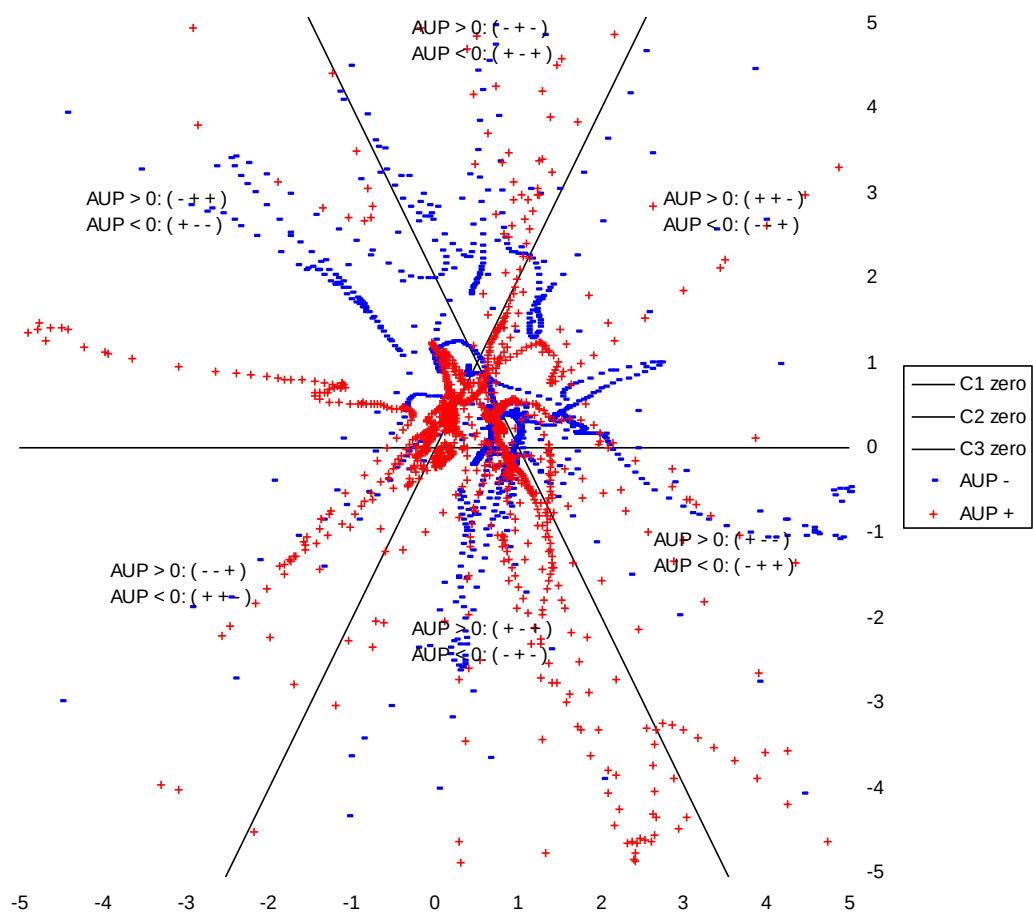


Figure 9.11: The Clustered Loadings L_{pxm} Showing the Latent Structure for the Filler Excipient Training Set shown in Table 9.6.

The loadings can then be mixed according to the ratios $\mathbf{R}_{\text{nxp}}$ to find the training set of scores $\mathbf{Q}_{\text{nxm}}$ as shown in Eq. 9.22. However, since each characterization experiment, particularly IR and NIR spectroscopy, are typically conducted only on pure components, Eq. 9.22 can be rewritten as Eq. 9.26.

$$Q_{iq} = \sum_{j=1}^p R_{ij} L_{jq}, \quad i \in u, \quad j \in p, \quad q \in m \quad (9.26)$$

This expression implies that the scores for each pure component are simply varying ratios of the underlying latent variable structure. Since the latent variable loadings are common across all molecules and molecular architecture in the training set, then the AUP^L in Eq. 9.24 can be used to create the score clusters, C_{iq}^Q . The AUP^Q of each pure component is found using Eq. 9.27.

$$AUP_i^Q = \sum_{j=1}^p AUP_j^L \cdot R_{ij} \quad (9.27)$$

The latent property lever arm, δ_{ij} , is then calculated using Eq. 9.28.

$$\delta_{ij} = \frac{R_{ij} \cdot AUP_j^L}{AUP_i^Q} \quad (9.28)$$

Finally the mixture score cluster for each $q \in m$ latent variable is calculated.

$$C_{iq}^Q = \sum_{j=1}^p \delta_{ij} \cdot C_{jq}^L \quad (9.29)$$

Conversely the score clusters can also be directly calculated from the scores using Eq. 9.30-9.32:

$$\Omega_{iq}^Q = \frac{Q_{iq}}{\psi_{ref,q}} \quad (9.30)$$

$$AUP_j^Q = \sum_{q=1}^m \Omega_{qj}^Q \quad (9.31)$$

$$C_{iq}^Q = \frac{\Omega_{iq}^Q}{AUP_j^Q} \quad (9.32)$$

Shown in Fig. 9.12 are conventional plots of the JMP scores given in Table 9.6 that illustrate the similarities in molecular architecture between excipients. As with the loadings plots, it is easier to distinguish contrasts in the molecular architecture in the plots that contain more descriptor information. Although the general influence of the loadings on the scores can be inferred by overlaying the loadings plots on the score plots, the magnitude of that influence is not directly apparent. In contrast, when the scores are clustered, the magnitude of each loadings impact on the score value can be directly calculated from the overlayed loading clusters. Fig. 9.11 shows the clustered JMP scores, which can be built through lever arm analysis from the scores listed in Fig. 9.9. Like the loading cluster diagram, two types of mixture score clusters are shown, those with *AUP* that are positive (circles) and those with *AUP* that are negative (squares). As demonstrated in Chapter 7, the clusters with negative *AUP* must be converted to clusters with positive *AUP* in order to properly interpret the solution. This can be accomplished by adjusting the reference values for the clusters using the algorithm developed in Chapter 6.

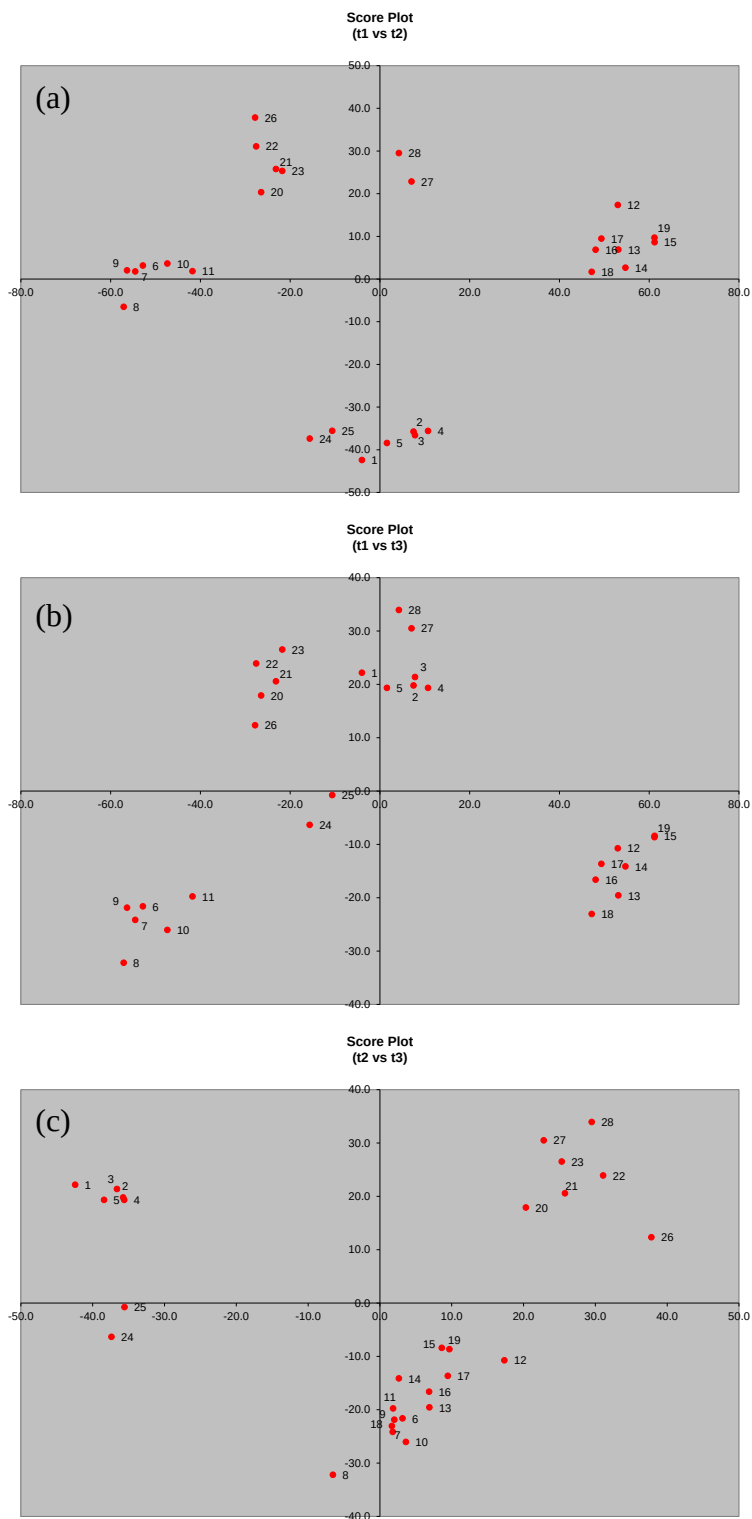


Fig. 9.12: From the PCA Decomposition of the Training Set Excipient Property Descriptors Shown in Fig. 9.8, Plots of the Loadings: (a) L_1 vs. L_2 , (b) L_1 vs. L_3 , and (c) L_2 vs. L_3 .

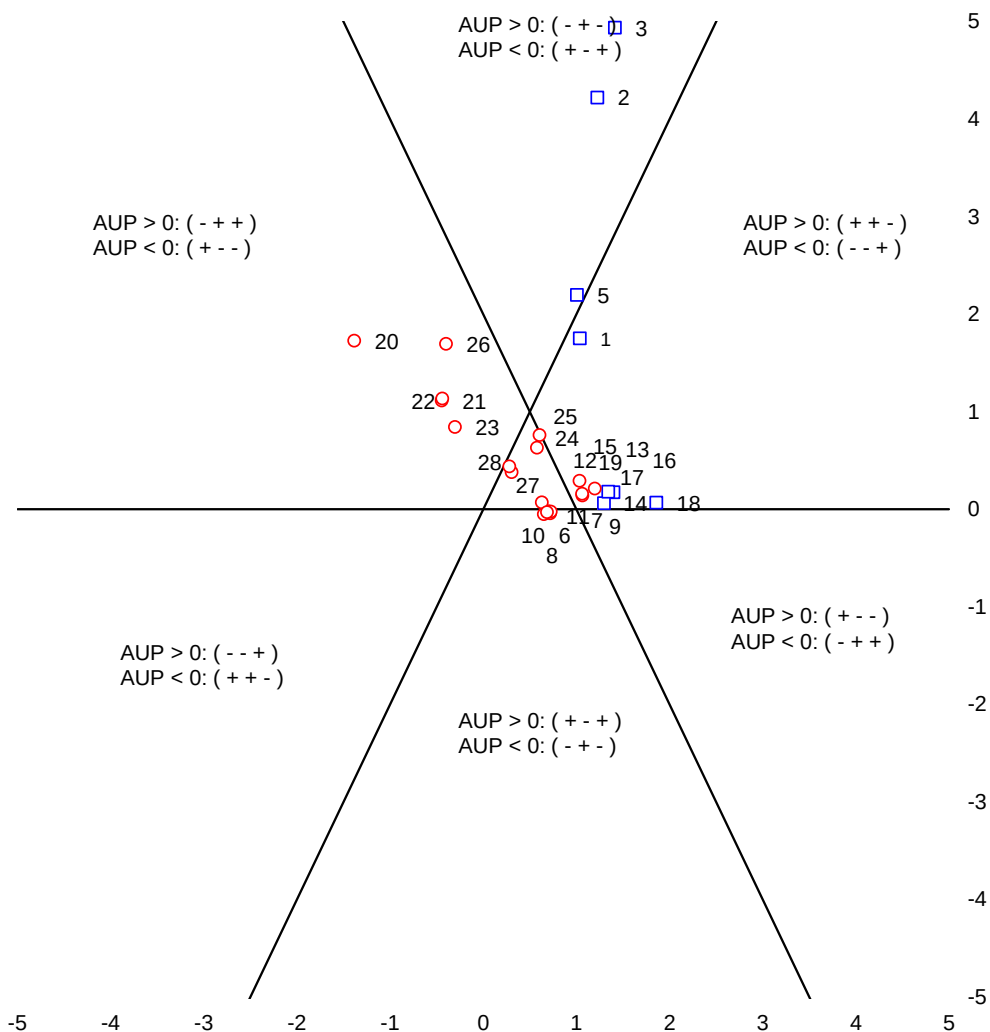


Figure 9.13: The Clustered Scores Q_{nxm} Showing the Latent Structure for the Filler Excipient Training Set Shown in Table 9.6.

By extension, since the score of a pure component can be constructed from ratios of its loadings, then the score of a molecular group can as well. Because the IR/NIR pure component descriptor data in the training set are a measure of the frequencies and intensities of group absorbances packaged together in a complete molecule, then the loadings can be treated as constant for any type of molecular architecture in the training set data. Hence, Eq. 9.26 can be rewritten as Eq. 9.33:

$$Q_{gq} = \sum_{j=1}^p R_{gj} L_{jq}, \quad g \in F, \quad j \in p, \quad q \in m \quad (9.33)$$

where the R_{gj} property ratios are estimated from the group specific IR/NIR responses generated by the procedure in Section 9.1.1. For example, to calculate the equivalent scores of $q=3$ latent properties for an CH_2 group, absorbances and intensities listed in Appendix A6 are converted to the IR response plot shown in Fig. 9.14.

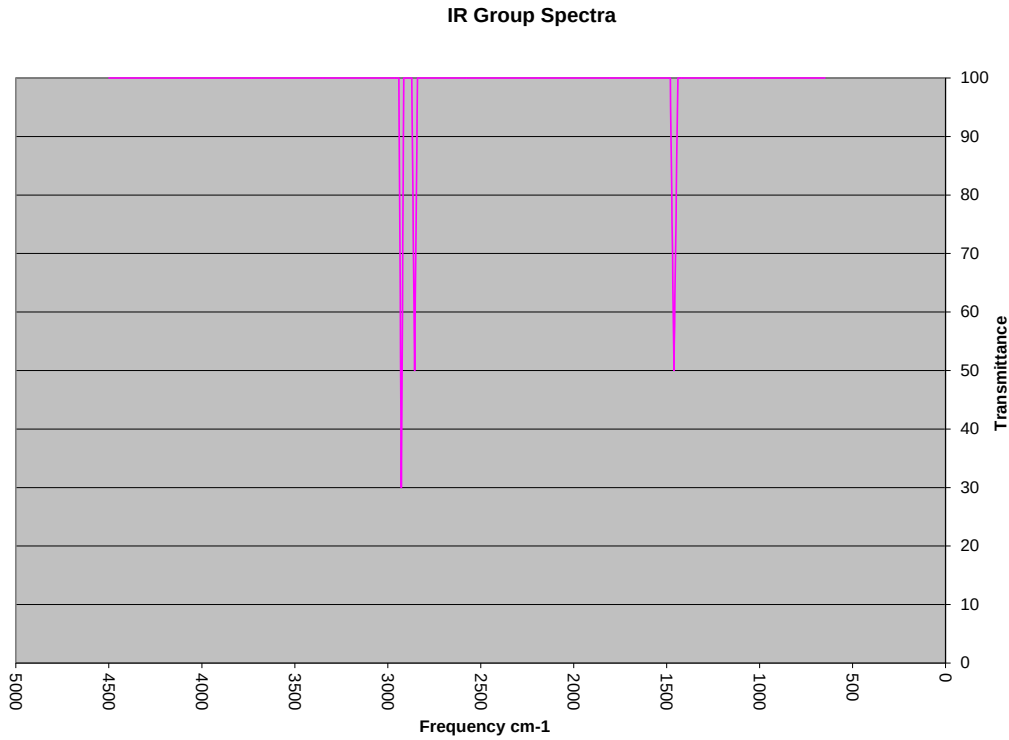


Figure 9.14: The IR transmittance response plot for a CH_2 group.

The group response data is then scaled and centered to generate $\mathbf{P}_{Fxp}$, followed by standardization using Eq. 9.34 and Eq. 9.35 to generate $\mathbf{R}_{Fxp}$:

$$S_g = \sum_j^p P_{jg}, \quad g \in F \quad (9.34)$$

$$R_{Exp} = (S_{F \times F})^{-1} P_{Exp} \quad (9.35)$$

The standardized Q_{Fxm} group scores can then be calculated using the matrix form of Eq. 9.33 shown in Eq. 9.36:

$$Q_{Fxm} = R_{Exp} L_{pxm} \quad (9.36)$$

To find the score clusters, C_{gq}^Q , the AUP of the score mixture is first calculated using Eq. 9.37, followed by the latent property lever arm, δ_{gj} , in Eq. 9.38.

$$AUP_g^Q = \sum_{j=1}^p AUP_j^L \cdot R_{gj} \quad (9.37)$$

$$\delta_{gj} = \frac{R_{gj} \cdot AUP_j^L}{AUP_g^Q} \quad (9.38)$$

Finally the group score cluster is calculated.

$$C_{gq}^Q = \sum_{j=1}^p \delta_{gj} \cdot C_{gq}^L \quad (9.39)$$

Conversely the score clusters can also be directly calculated from the scores using Eq. 9.40-9.42:

$$\Omega_{gq}^Q = \frac{Q_{gq}}{\psi_{ref,q}} \quad (9.40)$$

$$AUP_g^Q = \sum_{q=1}^m \Omega_{qg}^Q \quad (9.41)$$

$$C_{gq}^Q = \frac{\Omega_{gq}^Q}{AUP_g^Q} \quad (9.42)$$

The target feasibility region is constructed in the same manner. After converting a set of attribute targets into a set of T_{txq} latent property target scores (i.e. solving Eq. 9.13) using a procedure outlined in Chapter 8, the latent property targets are then recombined with the L_{pxm} loadings to generate a set of P_{txm} property descriptors of the property targets.

$$P_{txp} = T_{txm} L'_{mzp} \quad (9.43)$$

The scores are then standardized using Eq. 9.38 and 9.39.

$$S_t = \sum_j^p P_{jt} , \quad b \in t \quad (9.44)$$

$$Q_{txm} = (S_{txt})^{-1} T_{txm} \quad (9.45)$$

Conversely the score clusters can also be directly calculated from the scores using Eq. 9.46-9.48:

$$\Omega_{bq}^Q = \frac{Q_{gq}}{\psi_{ref,q}} \quad (9.46)$$

$$AUP_b^Q = \sum_{q=1}^m \Omega_{qb}^Q \quad (9.47)$$

$$C_{bq}^Q = \frac{\Omega_{bq}^Q}{AUP_b^Q} \quad (9.48)$$

A summary of the data generation and conversion process for cCAMD is shown in Fig. 9.15

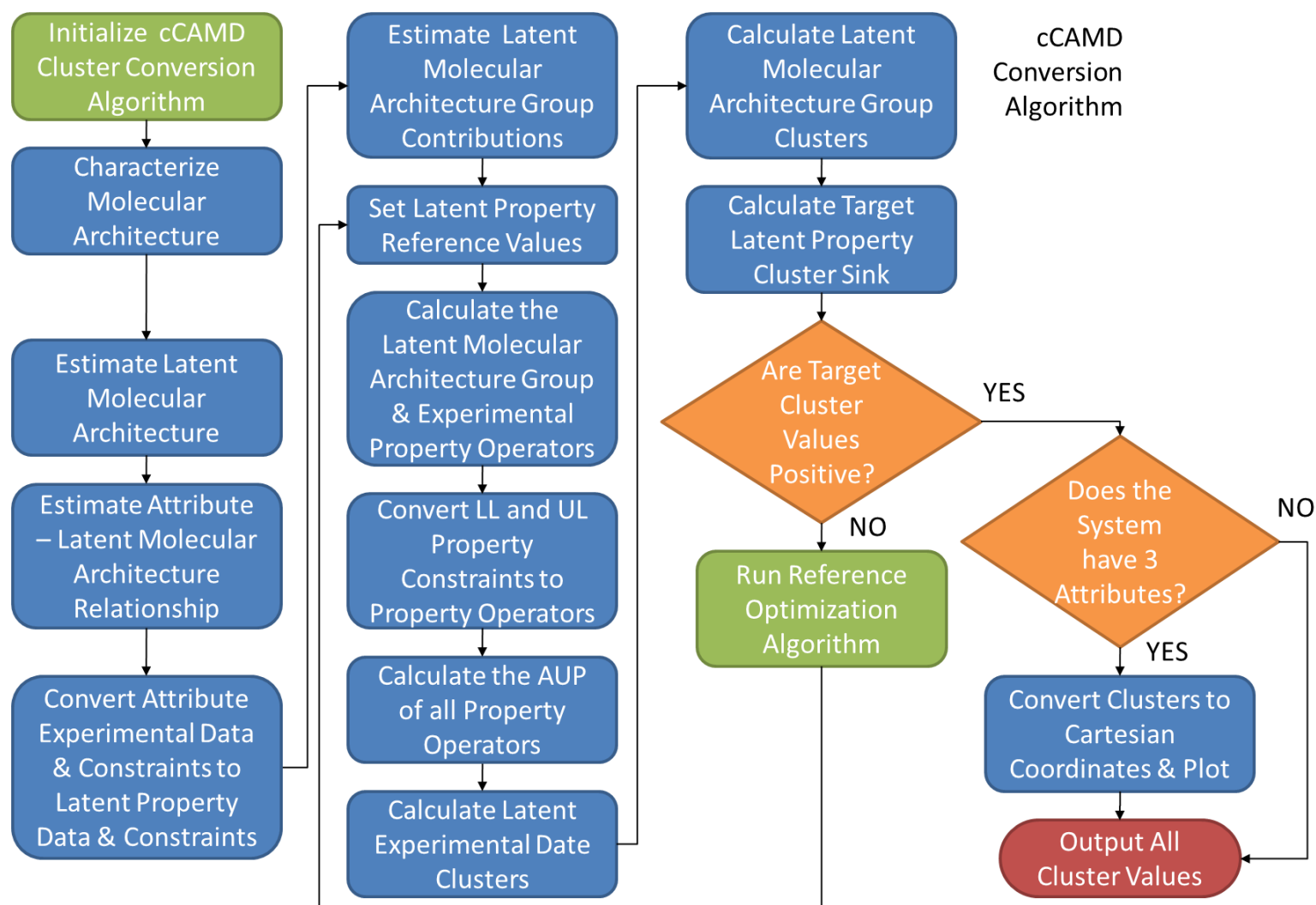


Figure 9.15: The cCAMD Cluster Conversion Flowchart.

9.3. cCAMD using Characterization Data and cGCM Models

It has been shown in Section 9.1 how characterization tools like IR and NIR spectroscopy can be used to describe molecular architecture and that this molecular architecture can be described using a new combinatorial tool called cGCM. It has also been shown how decomposition techniques can be used to compress this data structure to a more manageable arrangement suitable for multiscale chemical product design in Section 9.2. In order to conduct a molecular architecture design it is important to recognize that the constraints imposed by decomposition should also be observed for any new molecules or mixtures created. For instance, by definition, the mixture, pure component, and group representations of the underlying latent variable loadings are all linear mixture expressions.

$$T_{n \times m} = P_{n \times p} L_{p \times m} \quad (9.16)$$

In Eq. 3.43 it was shown how the latent variable scores can also be linearly mixed:

$$\hat{T}_M = XT \quad (3.43)$$

where X is a $u \times n$ matrix of chemical constituent mass fractions and T_M is a $n \times m$ matrix of candidate mixtures. Thus, candidate mixture scores can then be treated as *linear mixtures of linear mixtures* of the latent variable loadings. This observation assumes that any nonlinearity in the system is handled by the attribute-latent property relationship (Muteki *et al.* 2006). Violation of this constraint will significantly reduce the effectiveness of the cCAMD approach and can be avoided by ensuring the system is multivariate normal.

If Eq. 3.43 is set equal to the chemical product design target scores and solved for mass fractions of the chemical constituents in the training set, the procedure is referred to as a Case 1

mixture design in Section 3.2.4. In terms of property clusters, the mixture expression is first standardized using Eq. 9.45:

$$Q_{eq} = \sum_{i=1}^u Q_{iq} \cdot x_i \quad i \in u, \quad q \in m \quad (9.49)$$

where Q_{eq} is also known as Q_M or mixture score. The AUP , lever arm, and cluster expressions are then given by Eq. 9.50-9.52.

$$AUP_e^Q = \sum_{i=1}^u AUP_i^Q \cdot x_{ie} \quad (9.50)$$

$$\delta_{ie} = \frac{x_{ie} \cdot AUP_i^Q}{AUP_e^Q} \quad (9.51)$$

$$C_{eq}^Q = \sum_{i=1}^u \delta_{ie} \cdot C_{qi}^Q \quad (9.52)$$

Additionally, Garcia-Munoz *et al.* (2010) and Muteki and Macgegor (2006) have shown that the new mixture scores can be matched against a database to find new molecular architectures that are not part of the original training set, denoted as Case 2 in Chapter 3. However, the candidate molecular architectures found are subject to the completeness of the database. Using the cGCM property model on the group scores overcomes this shortcoming by providing a means to build all potential candidates that have the underlying latent variable loading structure of the training set. A set of group latent variables identified in the cGCM can then be combined to find new molecules that were not part of the original training set. The structure of the molecular design follows that of a first order GCM expression.

$$Q_{qi} = \sum_{g=1}^F Q_{qg} \cdot n_{ig} \quad (9.53)$$

The conversion of this approach to make use of the efficiency gains provided by the reverse problem formulation and property clustering results in a technique called characterization based computer aided molecular design (cCAMD) with clusters. The approach is similar the CAMD method discussed in Chapter 6, only it uses the latent property scores for the individual group contributions. Noting that the property operator expression of Eq. 9.53 can be non-dimensionalized, the expression becomes Eq. 9.54.

$$\Omega_{qi}^{MQ} = \sum_{g=1}^{n_g} n_g \Omega_{qg}^{MQ} \quad (9.54)$$

Rewriting the molecular property operator expression in terms of clusters is done by inserting Eq. 9.54 into Eq. 9.42 and rearranging to get Eq. 9.55-9.57.

$$C_{qi}^{MQ} = \frac{\sum_g n_g \cdot \Omega_{qg}^{MQ}}{AUP_i^{MQ}} = \frac{\sum_g n_g \cdot C_{qg}^{MQ} \cdot AUP_g^{MQ}}{AUP_i^{MQ}} \quad (9.55)$$

$$\delta_{g_1} = \frac{n_{g_1} \cdot AUP_{g_1}^{MQ}}{AUP_i^{MQ}} \quad (9.56)$$

$$C_{qi}^{MQ} = \sum_{g_1=1}^F \delta_{g_1} \cdot C_{qg_1}^{MQ} \quad (9.57)$$

The group based property model shown in Eq. 9.57 demonstrates how molecular groups or fragments can be added together analogous to inter-stream conservation (i.e. Eq. 9.52).

The method proceeds in a similar manner to that of CAMD, but with additional tests on the completeness of the molecular structures. Groups generated using the cCAMD algorithm are fed into the CAMD candidate generation algorithm. As with conventional CAMD, the maximum number of dissimilar groups μ_g^{max} is first set. Then, progressively larger combinations of dissimilar groups are added until the maximum is reached. The outputted solutions from the

CAMD algorithm are then tested sequentially against Rules 25-27. Any solutions that fail Rules 25-27 are discarded. Solutions that satisfy Rules 25-27 are deemed candidate solutions. The procedure is repeated until a maximum number of dissimilar groups μ_g^{max} is reached. A summary of the cCAMD candidate generation method is provided in Fig. 9.16.

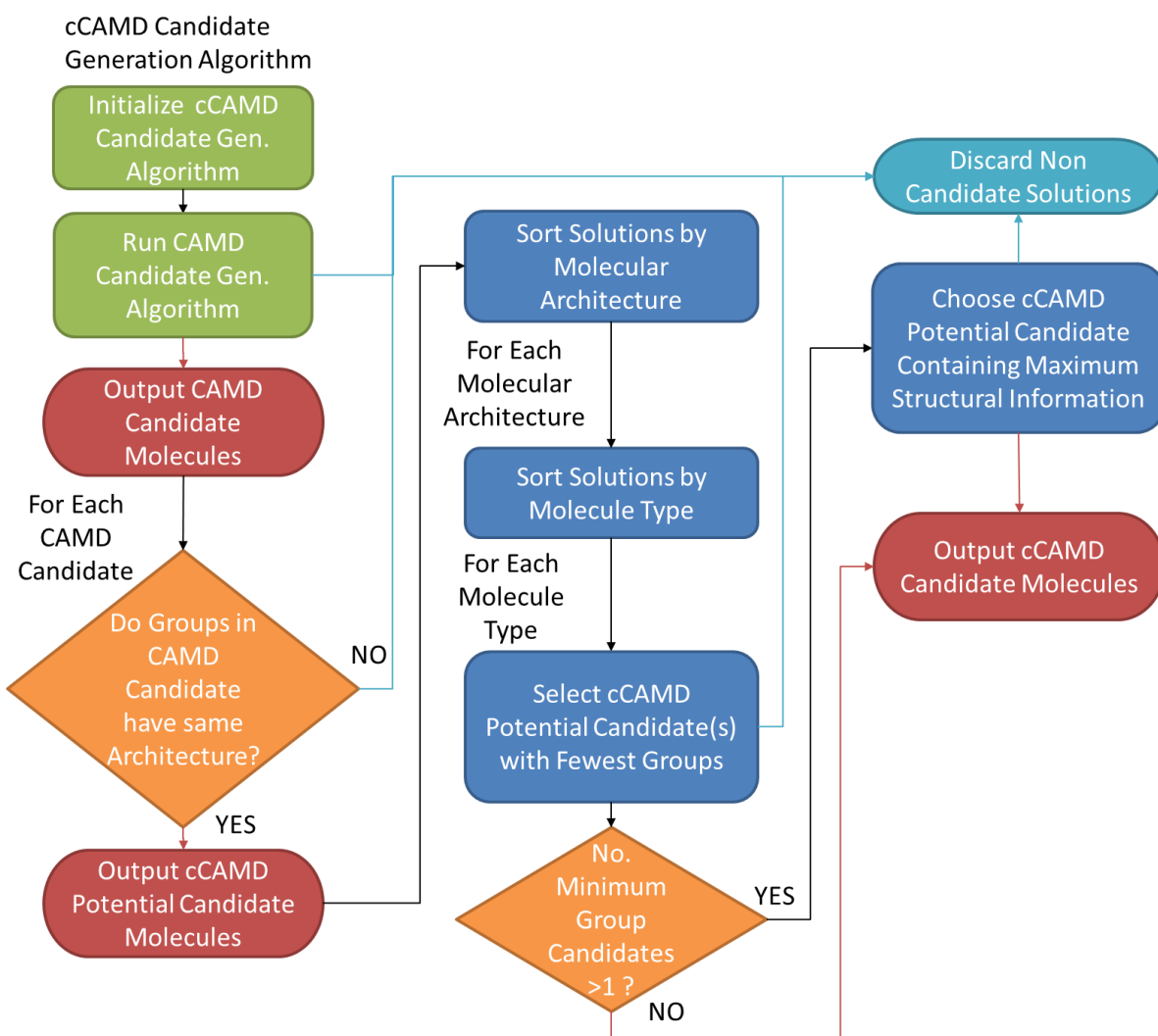


Figure 9.16: A Flowchart of the cCAMD Candidate Generation Algorithm using Clusters

It should be noted that the procedure calls the CAMD candidate generation procedure discussed in Table 6.5. For this instance, the expressions referenced in Table 6.5 have been replaced with the equivalent latent variable expressions, shown in Table 9.7.

Table 9.7: CAMD Candidate Generation Procedure with Latent Variables

Step	Description	Equation
1	Set Maximum Dissimilar Groups in the Mixture	-
2	Discard Pure Group Clusters (from CAMD Conversion Algorithm) that Fail Rule 1	6.7, 9.48, 9.32
3	Calculate No. Pure Groups	6.47
4	Calculate Pure Group Solutions, Discard Pure Group Solutions that Fail Rule 2 and Rule 3	6.7, 6.16, 6.42
5	Discard Pure Group Clusters that Fail Rule 4, Output Pure Group Candidate Molecules	6.41
6	For Binary Solutions with Pure Group Candidates, Calculate the No. Binary Groups	6.46, 6.48, 9.56
7	Calculate Solution AUP_M^M , Discard Solutions that Fail Rule 2	6.39, 6.47, 9.56
8	Calculate the FBN, Discard Solutions that Fail Rule 4	6.41
9	Calculate Solution Clusters, Discard Solutions that Fail Rule 1, and Output Candidate Solution Clusters	6.7, 9.32, 9.57
10	For Binary Solutions with no Pure Group Candidates, Calculate the Reference Constant for each Mixture	6.20
11	Calculate the Sink Constants for each Mixture, Discard Mixtures where all Sink Constants are $>$ or $<$ Reference Constant, Repeat Steps 6-9	6.23-6.24, 6.27-6.29
12	For Ternary and Larger Solutions ($\mu_g \geq 3$) containing $\mu_g - 1$ Candidates, Calculate the No. Groups, Repeat Steps 7-9	6.46, 6.48
13	For $\mu \geq 3$ Ternary+ Mixtures with no $\mu - 1$ Candidates, Calculate the Cluster Range of the Pure Components	6.34, 6.35, 6.6
14	Calculate the Cluster Range of the Feasibility Sink, Discard Mixtures whose Pure Component Cluster Range is $>$ Sink Cluster Range, Repeat Step 12	6.34, 6.35, 6.6, 6.7
15	Repeat Step 12-14 until μ_g^{max} is reached	-

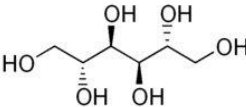
9.4. Case Study: Excipient Design for Direct Compressed Tablets

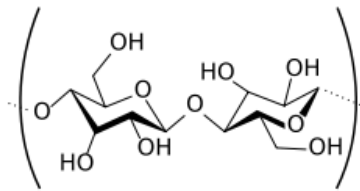
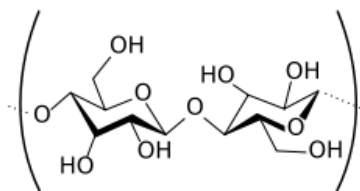
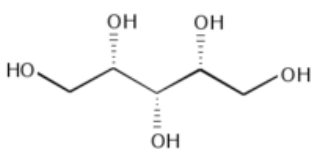
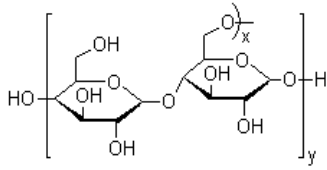
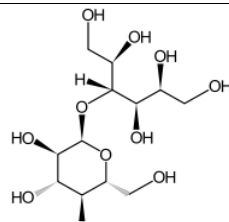
A detailed analysis of an excipient design of a direct compressed paracetamol tablet as discussed in Gabrielsson *et al.* (2002, 2003, 2004, 2006) serves as the primary case study for illustrating the approach presented in this chapter. A limited version of this case study is presented in this section as a proof-of-concept that demonstrates the method using components of the filler training set referenced in Section 9.1.

In Gabrielsson *et al.* (2003), the structure of a direct compressed paracetamol tablet is made up of three distinct classes of excipients (i.e. binders, disintegrants, and fillers) and the active ingredient acetaminophen. Binders generally hold the ingredients in a tablet together, providing mechanical strength. Disintegrants help the tablet expand and dissolve when entering the digestive tract. Fillers increase the bulk volume of the tablet, but can also serve in both a binding and disintegrant role as well. For this reason, the excipient design proposed by Gabrielsson *et al.* (2003) was further narrowed to focus on the design of only filler excipients.

A $n=24$ a list of the fillers for the proof of concept are shown in Table 9.8.

Table 9.8: Excipients used in the Proof-of-Concept Case Study

Fillers <i>i</i>	Common Name	Type of Filler	Molecular Structure	Particle Diameter (μm) and Structure*
1	Pearlitol SD200	Mannitol		200, g
2	Pearlitol 300DC			250, g
3	Pearlitol 400 DC			360, g
4	Pearlitol 500 DC			520, g
5	C Mannidex			N/A, g

6	Avicel PH-101	Microcrystalline Cellulose		50, c
7	Avicel PH-102			90, c
8	Avicel PH-112			90, c
9	Avicel PH-200	High Density Microcrystalline Cellulose		180, c
10	Avicel PH-301			50, c
11	Avicel PH-302			90, c
12	Xylitab DC	Xylitol		N/A, g
13	Xylisorb 90			90, c
14	Xylisorb 300			250, c
15	Xylisorb 700			700, c
16	Xylitab 100			200, g
17	Xylitab 200			300, g
18	Xylitab 300			150, g
19	C Xylidex CR 16055	Maltodextrin		200, g
20	Lycatab			300, g
21	C Sperse MD 01314			150, g
22	C Pharm 01980			N/A
23	C Pharm 01983			N/A
24	Maltisorb P90	Maltitol		N/A

Three representative property descriptors (i.e. P_1 , P_2 , and P_3) were chosen to evaluate the IR and NIR characterizon, as shown in Table 9.9.

Table 9.9: Property Descriptors of the Proof-of-Concept

Filler i	P_1	P_2	P_3
1	93	74	37
2	94	78	35
3	96	80	35
4	101	84	39
5	102	85	38
6	103	81	37
7	104	83	39
8	106	83	39
9	107	82	38
10	112	89	40
11	113	88	40
12	114	86	40
13	116	90	43
14	117	90	41
15	117	91	41
16	119	93	41
17	120	89	40
18	120	93	44
19	121	95	42

20	125	93	45
21	127	96	45
22	128	95	45
23	131	95	46
24	135	106	47

In order to ensure the system is multivariate normal, a natural log function was applied to each descriptor, and the system was centered, but not standardized to create $P_{n \times p}$. From this data set, a PCA decomposition was performed on the covariance matrix, which due to the extensive normalization, resulted in better description of the latent variable domain than that of a PCA decomposition on the correlation matrix. JMP 7.0.2 was chosen to perform the decomposition resulting in the $T_{n \times m}$ scores shown in Table 9.10.

Table 9.10: Latent Variable Scores of the Proof-of-Concept

Filler i	T_1	T_2	T_3
1	-0.2685	0.0611	0.0004
2	-0.2633	-0.0157	-0.0067
3	-0.2360	-0.0341	-0.0074
4	-0.1199	0.0141	-0.0367
5	-0.1207	-0.0150	-0.0286
6	-0.1526	-0.0089	0.0170
7	-0.1060	0.0166	-0.0084
8	-0.0930	0.0135	0.0052
9	-0.1064	-0.0012	0.0279
10	-0.0066	-0.0167	-0.0072

11	-0.0063	-0.0114	0.0062
12	-0.0120	0.0008	0.0268
13	0.0609	0.0280	-0.0126
14	0.0419	-0.0109	0.0090
15	0.0475	-0.0174	0.0021
16	0.0702	-0.0331	0.0007
17	0.0406	-0.0277	0.0420
18	0.1128	0.0213	-0.0163
19	0.1050	-0.0294	-0.0085
20	0.1524	0.0325	0.0055
21	0.1795	0.0111	-0.0029
22	0.1795	0.0161	0.0092
23	0.2068	0.0297	0.0186
24	0.2945	-0.0232	-0.0351

The L_{pxm} loadings associated with these scores are given in Table 9.11. For these results, it was determined that 98.5% of the response data could be explained using only the first two principal components (i.e. $m \approx 2$). This can be confirmed by investigating the scree plot shown in Fig. 9.17. However, for completeness, all three latent variables are used in the design.

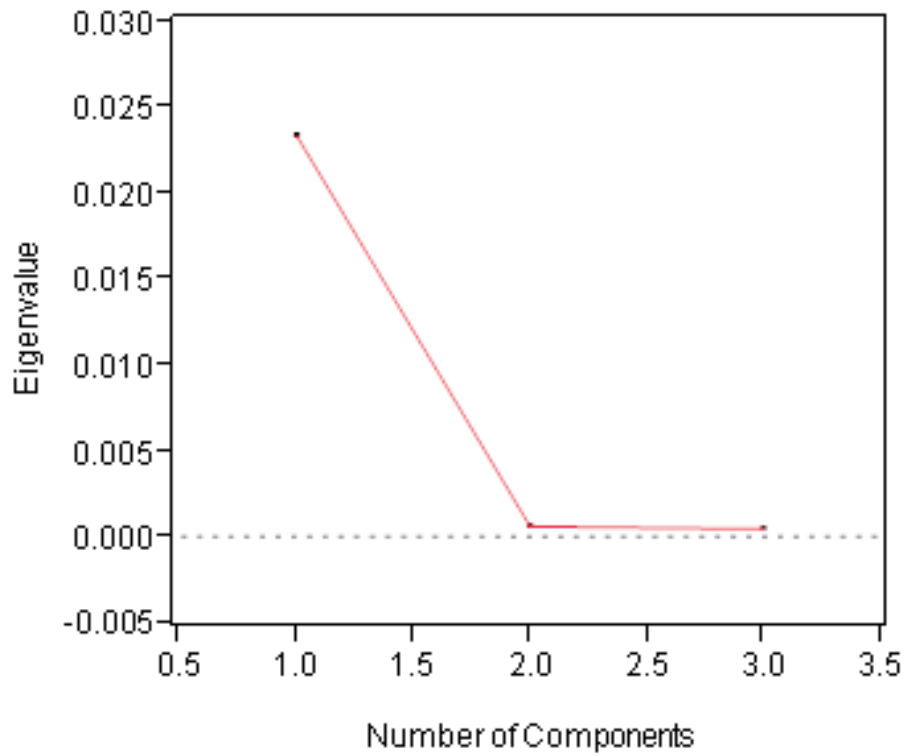


Figure 9.17: Scree plot of the Proof-of-Concept Principal Components

Table 9.11: Latent Variable Loadings of the Proof-of-Concept

Descriptor j	L_1	L_2	L_3
1	0.68310	-0.15948	0.71270
2	0.51022	-0.59401	-0.62195
3	0.52254	0.78849	-0.32440

Gabrielsson *et al* (2003) also identifies 8 different attributes important to the function of pharmaceutical excipients. These are as follows: disintegration time (DT), crushing strength (CS), Hausner ratio, disintegration type, ejection force (EF), adhesion, compression force, and tablet height. Disintegration time is the average time for a tablet to break down and pass through a sieve of a predetermined size. Crushing strength is the amount of force needed to break a tablet under compression and is a measure of a tablets strength and hardness. The Hausner ratio is a measure of how well a solid excipient powder flows. The disintegration type could be either bursting or erosion and could be used to describe the release profile. Ejection force is the strength needed to overcome the friction coefficient between the die wall and the tablet and is a function of adhesion. Adhesion is a measure of intermolecular bonding and microstructure. Compression force is the force used to compress the tablet to the appropriate tablet height. Of these attributes, disintegration time (DT), crushing strength (CS), and ejection force (EF) were selected for further study since these attributes offer a well-balanced description of the properties important to a pharmaceutical tablet.

Normally to develop the attribute-property relations (e.g. Eq. 9.13) a separate full factorial design of experiments (DOE) with centerpoints is executed where molecules containing the appropriate descriptor structure type are chosen to represent each of the design points. However, in this instance, attribute values were also measured for each chemical in the training set and shown in Table 9.12

Table 9.12: Design of Experiments Attribute Responses

Filler <i>i</i>	<i>DT</i>	<i>CS</i>	<i>EF</i>
1	0.0288	-1.4281	1.6503
2	0.0142	-1.8934	2.0880
3	0.0090	-1.8304	1.9957
4	0.0055	-0.5194	0.8334
5	0.0016	-0.7767	1.0158
6	0.0111	-1.2156	1.2004
7	0.0089	-0.5678	0.7014
8	0.0093	-0.5777	0.6168
9	0.0108	-0.9042	0.7995
10	-0.0034	-0.1222	0.1559
11	-0.0007	-0.1595	0.1154
12	0.0046	-0.2309	0.0747
13	-0.0010	0.6910	-0.6276
14	-0.0035	0.1613	-0.2528
15	-0.0059	0.1937	-0.2521
16	-0.0104	0.2479	-0.3261
17	-0.0018	-0.1553	-0.1527
18	-0.0062	1.0238	-0.9767
19	-0.0134	0.5674	-0.6080
20	-0.0039	1.2518	-1.3538
21	-0.0106	1.3364	-1.4229
22	-0.0081	1.3023	-1.4587
23	-0.0063	1.5332	-1.7526

24	-0.0287	2.0742	-2.0664

The three attributes *DT*, *CS*, and *EF* have been notoriously difficult to analyze using traditional mixing design because of the complex and highly nonlinear nature of pharmaceutical excipients (Martenello *et al.* 2008). In order to better control these attributes, they are mapped down to a domain subspace where they can be approximated as linear combinations of molecular and group parameters. The informative statistics of the attribute-latent property models (e.g. Eq. 9.13) are given in Table 9.13.

Table 9.13. Regression Coefficients for the Attribute-Latent Property Relationship

Informative Statistics	<i>DT</i>	<i>CS</i>	<i>EF</i>
β_1	-0.0676	6.9160	-7.5516
β_2	0.1731	7.0561	-6.1735
β_3	0.1371	-5.7299	-0.4081

For a good a molecular design it is desirable for the attributes to fall within the target ranges given in Table 9.14. Applying Eq. 9.13 to Table 9.14 results in the target scores shown in Table 9.15.

Table 9.14. Target Attributes for the Molecular Design

Attribute Targets	<i>DT</i>	<i>CS</i>	<i>EF</i>
<i>UL</i>	0.1909	1.764	-1.544
<i>LL</i>	-0.0169	1.729	-1.888

Table 9.15. Latent Variable Targets

Score Targets	T_1	T_2	T_3
<i>UL</i>	0.75	1.00	0.50
<i>LL</i>	0.25	0.00	0.00

These scores are then recombined with the loadings in Table 9.11 using Eq. 9.43 to generate the equivalent property descriptor bounds, corrected for log normal and centered behavior, as shown in Table 9.16.

Table 9.16: Predicted Property Descriptor Targets

Property Targets	P_1	P_2	P_3
<i>UL</i>	229.2	100.0	112.3
<i>LL</i>	229.2	52.2	46.2

In order to ensure that the domain is properly explored, the property responses at the six points that comprise the feasibility region (e.g. Eq. 6.7) are evaluated for the standardization matrix. The standardized latent variable targets are found by applying the standardization matrix to the latent variable scores at each of the six points and selecting the maximum and minimum resulting in Table 9.17.

Table 9.17. Attribute targets mapped to the descriptor sub-domain

Subspace Targets	Q_1	Q_2	Q_3	S
<i>UL</i>	0.801	2.881	1.602	1.321
<i>LL</i>	0.539	0.00	0.00	0.312

The molecular groups present in the training set and identifiable by the characterization were CH, CH₂, OH, CHO, O, CH₂-O, CHOH, CH₂OH, CHCH₂OH, CHCHO, α -pyranose, O_{cyc}, β -pyranose, and cellulose. The last two groups were only present in the micro-crystalline cellulose excipient and, as such, were excluded from the short chain molecular design. In the previous molecular design using these molecular groups with different target attributes (Solvason *et al.* 2009), the selection of the α stereoisomer of pyranose compounds as candidates was preferred, while the microcrystalline cellulose (MCC), which relies on the β stereoisomer was notably outside the target range (Solvason *et al.*, 2009). This was an unexpected result since MCC is not only one of the most common excipients used in acetaminophen tablets, but also contained the most information on long-range order when characterized. In order to better understand this conclusion, a simple investigation of the meso-scale microstructure was conducted by analyzing the influence of particle size (50-300 μ m) on the attribute properties.

There are three ways to illustrate the influence of particle size on the design problem. The first is to illustrate an average IR/NIR characterization for each particle size. Although a mathematically valid method for visualizing the particle size effect, a particle size ‘IR/NIR spectra’ has no physical meaning. A more appropriate method is to observe the effect of particle size on the individual chemical constituents in the model. Since increasing particle size results in larger reflective-absorbances (and smaller reflectances), the primary effect was an increase in the *AUP* values of the training set molecules and individual groups. A smaller, secondary effect due to a disproportionally strong effect observed in the 4500-6000 cm⁻¹ range for polyols (Storz and Steffens, 2004) caused a slight alteration of the location of the training set and molecular group clusters (Fig. 5.9). This influence suggests that a third method be employed, namely that functional groups be parameterized for different particle sizes and then recombined to build a list

of multiscale solutions simultaneously. Of the groups listed above, CHOH, CH₂OH, α -pyranose, β -pyranose, and cellulose exhibited changes in their property descriptors according to the diameter of the particle. For this proof-of-concept case study the final particle size should be in the 100-400 μm domain. Groups parameterized at 100 μm increments in this domain are shown in Table 9.18.

Table 9.18: Groups for Molecular Design of Excipient

Group g	Group Type	Particle Diameter (μm)	P_1	P_2	P_3
1	CH	-	85.5	52	18
2	CH ₂	-	0.5102902	60	15
3	OH	-	0.5225436	20	12
4	CHO	-	85.5	87	26.5
8	O	-	90	24	15
9	CH ₂ O	-	90	28	15
10	CHOH	100	63.1	46.2	23.3
11	CHOH	200	64.4	47.6	23.7
12	CHOH	300	65.6	49.0	24.1
13	CHOH	400	66.8	50.5	24.5
14	CH ₂ OH	100	78.9	106.2	26.3
15	CH ₂ OH	200	80.1	107.6	26.7
16	CH ₂ OH	300	81.3	109.0	27.1
17	CH ₂ OH	400	82.6	110.5	27.5
18	CHCH ₂ OH	-	81	160	27
19	CHCHO	-	135	60	30

20	Ocyc	-	72	16	18
21	α (pyranose)	100	67.2	78.9	31.6
22	α (pyranose)	200	67.2	79	31.5
23	α (pyranose)	300	67.3	79.1	31.4
24	α (pyranose)	400	67.4	79.2	31.3
25	β (pyranose)	100	109	141.8	31.8
26	β (pyranose)	200	111.5	144.3	34.3
27	β (pyranose)	300	114	146.8	36.8
28	β (pyranose)	400	116.5	149.3	39.3
29	Cellulose	100	84.7	103.4	42.2
30	Cellulose	200	82.6	100.6	41.4
31	Cellulose	300	80.4	97.8	40.6
32	Cellulose	400	78.2	95.0	39.7

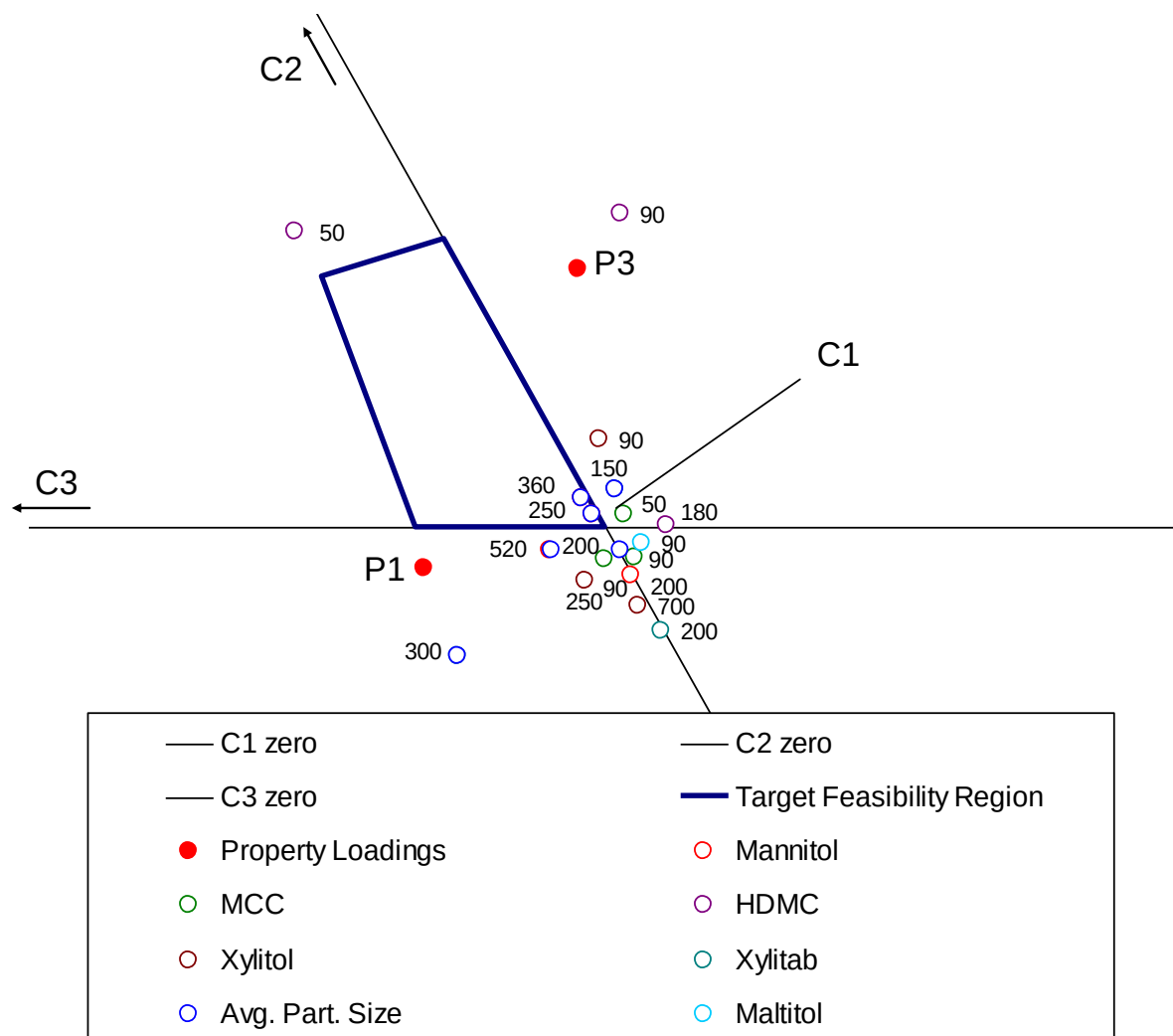


Figure 9.18. Design cluster diagram featuring training set particle sizes in μm .

Following the design algorithm presented in Section 9.3, an excipient design was conducted and the results are shown in Table 9.19. The designed structures are small diols, with one β -glucopyranose structure in the 200mm range. The β form of the glycosydic linkages are found in the most common excipient, microcrystalline cellulose. However, as the molecule size increases, aldehyde structures become more prevalent and the orientation preference in the glycosydic linkage disappears. In the candidate list published by Solvason *et al.* (2009) using slightly different constraints, the orientation preference is for the α -glucopyranose over the β -

glucopyranose structure. Rather, the indication is that particle size is the primary driving force in these structures as long as enough alcohol groups are present for sufficient intermolecular hydrogen bonding. This is a reasonable conclusion since ejection force, crushing strength, and disintegration time are primarily related to the strength of particle-particle interactions and only secondarily related to the molecular interactions in pharmaceutical powders.

Table 9.19. Designed Molecules and Discrete Particle Sizes of 100, 200, 300, and 400 μm .

Canidate Molecules	Q_1	Q_2	Q_3	Particle Size (μm)
CH-CH ₂ OH	1.19	0.74	0.33	300
CH ₂ OH-CH ₂ OH	1.26	1.52	0.91	300
CH ₂ OH-CH ₂ OH	1.25	1.41	0.83	200
CH ₂ OH- β (pyranose)-CH ₂ OH	1.75	0.05	0.00	200
OH-CH ₂ OH	1.19	0.68	0.29	200
CH ₂ OH-CH ₂ OH	1.24	1.31	0.77	100
OH-CH ₂ OH	1.19	0.63	0.25	100
CH ₂ OH-CH ₂ OH	1.27	1.66	1.00	400
CH ₂ OH-CHCHO-CH ₂ OH	1.72	1.72	0.09	400
OH-CH ₂ -O-CH ₂ OH	1.20	0.80	0.37	400

9.5. Summary

It was proposed in this chapter that a cCAMD method be developed such that attribute and property data are mapped into a property sub-domain, derived using decomposition techniques, that contains a better description of the effects of molecular architecture on the product's properties and attributes. The constraints on this new domain were that it should have a linear

property-molecular architecture relationship and that it should be capable of being described by combinatorial techniques in order to maximize the efficiency of the reverse problem formulation (RPF) with property clustering approach. A new method called characterization based group contribution method (cGCM) is proposed for this task and will result in molecular architecture information at multiple length-scales being utilized in the design. The approach also allowed the product designer the ability to customize the domain subspace to include only the molecular architecture information important to a specific set of product attributes and captured by characterization.

The combination of property clustering and principal component analysis offers many insights and advantages for the design of chemical products at multiple scales. In particular CAMD problems are no longer hindered by a lack of structure information in the molecular design. Rather, the uncertainty in predicting large molecular structures has now been removed from the models and replaced solely with the experimenter's ability to choose appropriate training sets, for which many proven techniques exist. This represents a useful addition to the existing CAMD methodology. Furthermore, the method is universal in nature and can be extended to include many characterization techniques.

Chapter 10

Conclusions

In this dissertation, a multi-scale product and process design framework based on the reverse problem formulation was developed for chemical products. The framework was designed to utilize experimental data, parameters, and models. The reason behind choosing to base the framework on experimental models comes from the nature of their intended end-use. Since the effectiveness of a chemical product was most often determined by its consumer attributes, and since the consumer attributes were most likely quantified using data from consumer preference tests, then it followed that the framework should at least be partially based on experimental data. This observation did not preclude using the more common physical and chemical properties found in process design in the framework, rather, it only means that experimental data needed to be integrated into the design in some fashion. Furthermore, while the use of data placed constraints on the application range of the tools developed for multiscale product design, it did not affect the framework developed and could be applied to systems described by deterministic relationships if desired. The use of experimental data provided insight into opportunities for improving the solution strategy for solving multiscale design problems.

10.1. Achievements

The first opportunity was the recognition that through the use of transformation functions and experimental design techniques, property and attribute models could be linearized. As a result, the duality of linear programming could be applied to solve the design problem in the lower dimensional property domain instead of the high dimensional component space. Approaching the design problem in this manner significantly reduced the computational complexity of the problem. Likewise, this opportunity was visualized in the property domain design space using property clustering, which provided insight into the design while reducing the computational complexity of the design problem. A new method called attribute based computer aided mixture design (aCAM^bD) with clusters was developed to take advantage of this opportunity for data-driven chemical product design problems and applied to a polymer blending problem. The method was built specifically to utilize data driven property models, regressor based property clusters, and the reverse problem formulation.

Recognizing that combinations of multiple sources of scientific knowledge, information, and data can be used in the design of chemical products, a second opportunity was identified. In particular, it was observed that the effectiveness of most chemical products is determined by its consumer attributes which did not have QSPR/QSAR relationships. Alternatively, the attribute information could be mapped into a physical property domain with strong QSAR/QSPR descriptions. As a result, a second new method called (aCAMD) with clusters was developed to map consumer attribute data into a more conventional physical property domain where GCM was used to find new chemical structures. The method was built to use data driven attribute-property relationships, GCM based property models to design molecules, and property clusters in

a reverse problem formulation to reduce the computational complexity of the problem. The method was applied in the design of environmentally benign refrigerants.

A third opportunity resulted from the recognition that the various sources of information used in multiscale chemical product design come from different length scales, each with particular molecular architectures that dominate the chemistry at that scale. Quantification of these sources of information would be best achieved through the use of characterization techniques which could be combined to handle complex molecular architecture descriptions. Combining the different sets of characterization information would result in a large, highly correlated data set that completely describes the molecular architecture of the chemical product. Rather than solving the molecular design in this high dimensional data set, decomposition techniques could be applied to the characterization data set to find a lower dimensional latent variable structure. The relationship between the underlying latent variable structure and the consumer attributes could then be estimated using chemometric techniques. In addition, a new characterization based group contribution method (cGCM) could be developed to take advantage of the additional molecular architecture information provided by characterization. As a result, a third new method called characterization based computer aided molecular design (cCAMD) was developed to map consumer attribute data to a latent variable domain where a newly developed cGCM method was used to find candidate molecular architectures. In particular characterization information from the molecular scale on short range order, such as group structure, conformation, and stereoregularity, can be combined with information from the micro-scale on long range order, such as the size, shape, and aspect ratio of particles within a solid matrix. The method was built specifically to utilize IR/NIR spectroscopy data, PCA decomposition techniques, MLR attribute-latent variable relationships, latent variable property clusters, and the

reverse problem formulation. The method results in a framework that decouples multiscale design problems from their respective scales, solving them in a low dimensional, computationally efficient manner. The method was applied to a constrained pharmaceutical excipient design with only three wavelengths.

10.2. Challenges and Future Work

Immediate extensions of the work presented in this dissertation include the use of partial least squares on to latent surfaces (PLS) to simultaneously develop the attribute-latent property and property descriptor – latent property relationships. This technique sacrifices some descriptor predictive capability for improvements in the attribute-latent property predictive power. Case studies incorporating the design of blend fractions with a predefined molecular architecture under process constraints could be used to both highlight the PLS improvement, as well as address the Case 3 class of problems discussed in Chapter 3. The same approach could also be used to select and design molecular architectures and process parameters to address Case 4 classes of problems.

One of the constraints of the property clustering method using decomposition techniques is that the developed characterization based group contribution method (cGCM) has been constructed to utilize only first order groups. It can be shown that this method can be extended to second groups which open the method to more common chemical engineering properties like binary interaction parameters that are common in many equations of state. Use of the reverse problem formulation and the developed property clustering methods, augmented for binary interaction parameters, would help to solve flash calculations and separations of unknown mixture fractions quickly. In particular, it is the author's belief that this can be accomplished by treating binary systems as systems of ternary mixtures where the third component is a binary

interaction parameter. The lever arms associated with the binary interaction parameter has been shown to be a function of the normal lever arm between the two components. Hence, at worst an iterative scheme could be developed to back calculate the lever arm at a design point.

Long term, the Euclidean vector suggests that linear mixing can extend to multiple properties so long as the conversion between clusters and property axes are defined. In other words, all cluster design problems should be capable of being described by Euclidean vectors. This opens up the design problems to handle an increasing number of molecular architecture and characterization techniques not explored in this dissertation. Likewise, since computational chemistry techniques are capable of generating characterization parameters, it should be possible to utilize the reverse problem formulation to redefine the manner in which computational chemistry is performed. In particular, spectroscopy has been shown to follow time dependent computational chemistry techniques that make use of the Hamiltonian functions. Methods to express the eigenvalue functions of the Hamiltonian in algebraic terms could open up computational chemistry techniques for use in chemical engineering product and process design.

References

- Abebe, S. B., X. Z. Wang, R. Li, K. J. Roberts and X. Lai (2008). "The information content in NIR spectral data for slurries of organic crystals." *Powder Technology* (179): 176-183.
- Achenie, L. E. K., R. Gani and V. Venkatasubramanian, Eds. (2003). *Computer Aided Molecular Design: Theory and Practice*. Computer Aided Chemical Engineering, 12, Elsevier.
- Achenie, L. E. K. and M. Sinha (2004). "The design of blanket wash solvents with environmental considerations." *Advances in Environmental Research* 8(2): 213-227.
- Akkisetty, P. V., G. V. Reklaitis, V. Venkatasubramanian (2009). "Ontological informatics based decision support system for pharmaceutical product development: Milling as a case study." *Computer Aided Chemical Engineering*, 26, 159-164.
- Alexander, L. and H. P. Klug (1948). "Basic aspects of X-ray absorption in quantitative diffraction analysis of powder mixtures." *Analytical Chemistry* 20(9): 789-795.
- Almaya, A. and A. Aburub (2008). "Effect of Particle Size on Compaction of Materials with Different Deformation Mechanisms with and without Lubricants." *AAPS PharmaSciTech* 9(2): 414-418.
- Almutlaq, A. M., V. Kazantzi and M. M. El-Halwagi (2005). "An algebraic approach to targeting waste discharge and impure fresh usage via material recycle/reuse networks." *Clean Technologies and Environmental Policy*: 1-12.
- Ambrose, D. (1978). "Correlation and estimation of vapor-liquid critical properties: 1. Critical temperatures of organic compounds." NPL Report Chem (United Kingdom, National Physical Laboratory, Division of Chemical Standards) 92.
- Ananda, J. and G. Herath (2009). "A critical review of multi-criteria decision making methods with special reference to forest management and planning." *Ecological Economics* 68(10): 2535.
- Anastas, P. T. and J. C. Warner (1998). *Green Chemistry: Theory and Practice*. New York, NY, Oxford University Press.
- Anderson, A. (2009). *Designer Silica Layers for Advanced Applications: Processing and Properties*. Dept. of Chemical Engineering. Auburn, AL, Auburn University. PhD: 191.
- Ansmann, A., H. Gondek, R. Kawa and H. Tesmann (2001). *Sun screen agents*. U. P. Office. USA. 6280712.

- Atkins, P. (1998). *Physical Chemistry*. New York, NY, W. H. Freeman and Company.
- Baskin, I. I., E. V. Gordeeva, R. O. Devdariani, N. S. Zefirov, V. A. Palyulin and M. I. Stankevich (1990). "Methodology of Solution of the Inverse Problem for the Structure-Property Relationship for the case of Topological Indices." *Doklady Akademii Nauk SSSR: Chemistry (English Translation)* 307: 217.
- Baughman, R. H., A. A. Zakhidov and W. A. de Heer (2002). "Carbon Nanotubes--the Route Toward Applications." *Science* 297(5582): 787-792.
- Bicerano, J. (1996). *Prediction of Polymer Properties*. New York, NY, Marcel Dekker, Inc.
- Bie, H. d. (2010). *ACD / Labs: Chemical structure drawing, Nomenclature, Analytical Data Processing and Management, Computer Assisted Structure Elucidation, Physicochemical Property Prediction*.
- Biegler, L. T. and I. E. Grossmann (1997). *Systematic Methods of Chemical Process Design*. Upper Saddle River, NJ, Prentice Hall PTR.
- Borchers, B. (1992). "A primal-dual interior point cutting plane method for the linear ordering problem." *COAL Bulletin* 21: 13-18.
- Borchers, B. and J. E. Mitchell (1994). "An improved branch and bound algorithm for mixed integer nonlinear programs." *Comput. Oper. Res.* 21(4): 359-367.
- Bovey, F. (1982). *Chain Structure and Conformation of Macromolecules*, Academic Press.
- Bovey, F. (1988). *Nuclear Magnetic Resonance Spectroscopy (2nd Ed.)*, Academic Press.
- Box, G., W. Hunter and J. Hunter (1978). *Statistics for Experimenters*. New York, NY, John Wiley & Sons.
- Boznak, R. G. (1993). *Competitive Product Development*. Milwaukee, WI, One Irwin/Quality Press.
- Brandvik, P. J. (1998). "Statistical simulation as an effective tool to evaluate and illustrate the advantage of experimental designs and response surface methods." *Chemometrics and Intelligent Laboratory Systems* (42): 51-61.
- Brandvik, P. J. and P. S. Daling (1998). "Optimisation of oil spill dispersant composition by mixture design and response surface methods." *Chemometrics and Intelligent Laboratory Systems* (42): 63-72.
- Brignole, E. A. and M. Cismonti (2003). *Molecular Design - Generation & Test Methods. Computer Aided Molecular Design: Theory and Practice*. L. E. K. Achenie, R. Gani and V. Venkatasubramanian, Elsevier. 12: 23-41.

Broadbelt, L. J., C. S. Henry and V. Hatzimanikatis (2009). "Discovery of novel routes for the production of fuels and chemicals." *Design for Energy and Environment*, 10, 141-147.

Brown, J. (2007). "Predicting performance of refrigerants using the Peng-Robinson Equation of State." *International Journal of Refrigeration*(30): 1319-1328.

Buback, M. and A. A. Harfoush (1983). "Near-infrared absorption of pure n-heptane between 5000 cm⁻¹ and 6500 cm⁻¹ to high pressures and temperatures." *Z. Naturforsch* 38a: 528-532.

Buback, M. and H. P. Voegelé (1993). *FT-NIR Atlas*. New York, NY, VCH.

Burns, D. A. and E. W. Ciurczak (2008). *Handbook of Near-Infrared Analysis*. New York, NY, CRC Press.

Cabezas, H. (2000). "Designing green solvents." *Chemical Engineering*: 107-109.

Camarda, K. V. and C. D. Maranas (1999). "Optimization in Polymer Design Using Connectivity Indices." *Industrial & Engineering Chemistry Research* 38(5): 1884-1892.

Caravajal, G. S., D. E. Leyden, G. R. Quinting and G. E. Maciel (1988). "Structural characterization of (3-aminopropyl)triethoxysilane-modified silicas by silicon-29 and carbon-13 nuclear magnetic resonance." *Analytical Chemistry* 60(17): 1776.

Chalivendra, V. B., S. Hong, I. Arias, J. Knap, A. Rosakis and M. Ortiz (2009). "Experimental validation of large-scale simulations of dynamic fracture along weak planes." *International Journal of Impact Engineering* 36(7): 888.

Chambost, V., R. Eamer and P. R. Stuart (2007). *Systematic Methodology for Identifying Promising Forest Biorefinery Products*. 93rd Pulp and Paper Technical Association of Canada (PAPTAC), Montreal, QC.

Chang, A.-F. and Y. A. Liu (2010). "Integrated Process Modeling and Product Design of Biodiesel Manufacturing." *Industrial & Engineering Chemistry Research* 49(3): 1197.

Charpentier, J.-C. (2002). "The triplet "molecular processes-product-process" engineering: the future of chemical engineering?" *Chemical Engineering Science* 57(22-23): 4667.

Charpentier, J.-C. (2009). "Perspective on multiscale methodology for product design and engineering." *Computers & Chemical Engineering* 33(5): 936.

Chemangattuvalappil, N. G., F. T. Eljack, C. C. Solvason and M. R. Eden (2009). "A novel algorithm for molecular synthesis using enhanced property operators." *Computers & Chemical Engineering* 33(3): 636-643.

Chemangattuvalappil, N. G., C. C. Solvason, S. Bommareddy and M. R. Eden (2010). "Combined property clustering and GC+ techniques for process and product design." *Computers & Chemical Engineering* 34: 582-591.

Chemmanagattuvalappil, N. G., C. C. Solvason, S. Bommareddy and M. R. Eden (2010). "Molecular Signature Descriptors for Integrated Flowsheet and Molecular Design." *Computer Aided Chemical Engineering* 28: 1267-1272.

Chemmanagattuvalappil, N. G., C. C. Solvason and M. R. Eden (2009). "Property based product design using combined property clustering and GC+ techniques." *Computer-Aided Chemical Engineering* 26: 237-242.

Cheng, H. N. and A. D. English (2003). *NMR Spectroscopy of Polymers in Solution and in the Solid State*, Oxford University Press.

Christensen, C. (1997). *The Innovator's Dilemma*, Harvard Business School Press.

Churchwell, C. J., M. D. Rintoul, S. Martin, D. P. Visco, A. Kotu, R. S. Larson, L. O. Sillerud, D. C. Brown and J.-L. Faulon (2004). "The signature molecular descriptor 3. Inverse-quantitative structure-activity relationship of ICAM-1 inhibitory peptides." *Journal of Molecular Graphics & Modelling* 22(4): 263-273.

Ciurczak, E. (1987). "Uses of Near-Infrared Spectroscopy in Pharmaceutical Analysis." *Applied Spectroscopy Reviews* 23(1): 147 - 163.

Committee on Integrated Computational Materials Engineering, N. R. C. (2008). *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security*, The National Academies Press.

Coniglio, L. and J. L. Daridon (1997). "A group contribution method for estimating ideal gas heat capacities of hydrocarbons." *Fluid Phase Equilibria* 139(1-2): 15-35.

Connolly, J. R. (2007). "Elementary Crystallography for X-Ray Diffraction, prepared for EPS400-001." Retrieved 8/21/2009.

Connolly, J. R. (2007). "Introduction to X-ray Powder Diffraction, prepared for EPS400-002." Retrieved 8/19/2010, from <http://epswww.unm.edu/xrd/xrdclass/01-XRD%20Intro-PPT.pdf>.

Connolly, J. R. (2009). "Diffraction Basics Part 2, prepared for EPS400-01." Retrieved 8/21/2010, from <http://epswww.unm.edu/xrd/xrdclass/06-Diffraction-II.pdf>.

Connolly, J. R. (2009). "Diffraction Basics, prepared for EPS400-001." Retrieved 8/21/2010, from <http://epswww.unm.edu/xrd/xrdclass/05-Diffraction-Basics.pdf>.

Constantinou, L., K. Bagherpour, R. Gani, J. A. Klein and D. T. Wu (1996). "Computer aided product design: problem formulations, methodology and applications." *Computers & Chemical Engineering* 20(6-7): 685-702.

Constantinou, L. and R. Gani (1994). "New group contribution method for estimating properties of pure compounds." *AIChE Journal* 40(10): 1697-1710.

Constantinou, L., R. Gani and J. P. O'Connell (1995). "Estimation of the acentric factor and the liquid molar volume at 298 K using a new group contribution method." *Fluid Phase Equilibria* 103(1): 11-22.

Constantinou, L., S. E. Prickett and M. L. Mavrovouniotis (1993). "Estimation of thermodynamic and physical properties of acyclic hydrocarbons using the ABC approach and conjugation operators." *Industrial & Engineering Chemistry Research* 32(8): 1734-46.

Conte, E., A. Martinho, H. A. Matos and R. Gani (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.

Conte, E., R. Morales-Rodriguez, R. Gani and C. A. O. d. N. a. E. C. B. Rita Maria de Brito Alves, Jr. (2009). *The Virtual Product-Process Design Laboratory for Design and Analysis of Formulations*. Computer Aided Chemical Engineering, Elsevier. Volume 27: 825.

Cooper, R. G. and S. J. Edgett (2009). *Product Innovation and Technology Strategy*. United States, Product Development Institute.

Corey, E. J. and W. T. Wipke (1969). "Computer-Assisted Design of Complex Organic Syntheses." *Science* 166(3902): 178-192.

Cornell, J. A. (2002). *Experiments with Mixtures*. New York, NY, John Wiley & Sons.

Costa, R., G. D. Moggridge and P. M. Saraiva (2006). "Chemical product engineering: An emerging paradigm within chemical engineering." *AIChE Journal* 52(6): 1976-1986.

Cox, D. R. (1971). "A note on polynomial response functions for mixtures." *Biometrika* 58(1): 155-159.

Cussler, E. L. and G. D. Moggridge (2001). *Chemical Product Design*. New York, Cambridge University Press.

Cussler, E. L., A. Wagner and L. Marchal-Heussler (2010). "Designing chemical products requires more knowledge of perception." *AIChE Journal* 56(2): 283-288.

d'Anterrosches, L. (2010). *Cheme'o-High Quality Chemical Properties*.

d'Anterrosches, L. and R. Gani (2004). "Group contribution based process flowsheet synthesis, design and modelling." *Computer-Aided Chemical Engineering* 18: 379-384.

d'Anterrosches, L. and R. Gani (2005). "Group contribution based process flowsheet synthesis, design and modelling." *Fluid Phase Equilibria* 228-229: 141-146.

d'Anterrosches, L. and R. Gani (2005). "Novel group contribution based approach in process flowsheet synthesis, design and modelling." *Computer-Aided Chemical Engineering* 20A: 643-648.

d'Anterroches, L. and R. Gani (2005). "The reverse approach for synthesis and design of chemical products & processes." World Congress of Chemical Engineering, 7th, Glasgow, United Kingdom, July 10-14, 2005: 86912/1-86912/9.

d'Anterroches, L. and R. Gani (2006). "Computer aided methodology for simultaneous synthesis, design & analysis of chemical products processes." Computer-Aided Chemical Engineering 21A: 853-858.

d'Anterroches, L., R. Gani, P. M. Harper and M. Hostrup (2005). "Design of molecules, mixtures and processes through a novel group contribution method." World Congress of Chemical Engineering, 7th, Glasgow, United Kingdom, July 10-14, 2005: 86916/1-86916/10.

Dai, J., L. Jin and L. Wang (1998). "Prediction of molar volume of aliphatic compounds using edge adjacency index." Progress in Natural Science 8(6): 754-761.

Daudel, R., R. Lefebvre and C. Moser (1959). Quantum chemistry, methods and applications. New York, Interscience publishers, inc.

Deeley, C. M. and R. A. Spragg (1991). "A comparison of Fourier transform infrared and near-infrared Fourier transform Raman spectroscopy for quantitative measurements: an application in polymorphism." Spectrochim. Acta. 47A(9/10): 1217-1223.

Desiraju, G. (1996). "The CHO Hydrogen Bond: Structural Implications and Supramolecular Design." Accounts of Chemical Research (29): 441-449.

Diwekar, U. M. (2003). "Introduction to Applied Optimization."

Doherty, M. F. and D. A. Green (2004). Crystal Engineering for Product and Process Design. Foundations of Computer-Aided Process Design. C. A. Floudas and R. Agrawal: 149-163.

Dunn, R. F. and M. M. El-Halwagi (1994). "Optimal design of multicomponent VOC condensation systems." Journal of Hazardous Materials 38(1): 187-206.

Duvedi, A. P. and L. E. K. Achenie (1996). "Designing environmentally safe refrigerants using mathematical programming." Chemical Engineering Science 51(15): 3727.

Dym, C. L. and P. Little (2008). Engineering Design: A Project Based Introduction. New York, John Wiley & Sons.

Eden, M. R., R. Gani and C. C. Solvason (2010). Process and Product Synthesis/Design: Product Design. Ullman's Encyclopedia of Industrial Chemistry. R. Gani, Wiley-VCH (accepted for publication).

Eden, M. R., S. B. Jorgensen, R. Gani and M. El-Halwagi (2002). "Property integration - a new approach for simultaneous solution of process and molecular design problems." Computer-Aided Chemical Engineering 10: 79-84.

Eden, M. R., S. B. Jorgensen, R. Gani and M. M. El-Halwagi (2003). "Property cluster based visual technique for synthesis and design of formulations." *Computer-Aided Chemical Engineering* 15B: 1175-1180.

Eden, M. R., S. B. Jorgensen, R. Gani and M. M. El-Halwagi (2004). "A novel framework for simultaneous separation process and product design." *Chemical Engineering and Processing* 43(5): 595-608.

Eden, M. R., S. B. Jørgensen, R. Gani and M. M. El-Halwagi (2003). "Reverse Problem Formulation based Techniques for Process and Product Design." *Computer Aided Chemical Engineering* 15A: 451-456.

Edwards, S. A. (2007). *The Promise of Nanotechnology. The Nanotech Pioneers*: 1-14.

El-Halwagi, M. (1997). *Pollution Prevention Through Process Integration: Systematic Design Tools*. San Diego, CA, Academic Press.

El-Halwagi, M. (2006). *Process Integration. Process Systems Engineering*. G. Stephanopoulos and E. N. Pistikopoulos. Amsterdam, Academic Press. 7.

El-Halwagi, M. M. (1992). "Synthesis of reverse-osmosis networks for waste reduction." *AIChE Journal* 38(8): 1185-1198.

El-Halwagi, M. M., I. M. Glasgow, X. Qin and M. R. Eden (2004). "Property integration: Componentless design techniques and visualization tools." *AIChE Journal* 50(8): 1854-1869.

El-Halwagi, M. M., A. A. Hamad and G. W. Garrison (1996). "Synthesis of waste interception and allocation networks." *AIChE Journal* 42(11): 3087-3101.

El-Halwagi, M. M. and V. Manousiouthakis (1989). "Synthesis of mass exchange networks." *AIChE Journal* 35(8): 1233-1244.

El-Halwagi, M. M. and V. Manousiouthakis (1990). "Automatic synthesis of mass-exchange networks with single-component targets." *Chemical Engineering Science* 45(9): 2813-2831.

El-Halwagi, M. M. and V. Manousiouthakis (1990). "Simultaneous synthesis of mass-exchange and regeneration networks." *AIChE Journal* 36(8): 1209-1219.

El-Halwagi, M. M. and B. K. Srinivas (1992). "Synthesis of reactive mass-exchange networks." *Chemical Engineering Science* 47(8): 2113-2119.

El-Halwagi, M. M., B. K. Srinivas and R. F. Dunn (1995). "Synthesis of optimal heat-induced separation networks." *Chemical Engineering Science* 50(1): 81-97.

Eljack, F. T., A. F. Abdelhady, M. R. Eden, F. B. Gabriel, X. Qin and M. M. El-Halwagi (2005). "Targeting optimum resource allocation using reverse problem formulations and property clustering techniques." *Computers & Chemical Engineering* 29(11-12): 2304-2317.

Eljack, F. T., R. M. Cummings, A. F. Abdelhady, M. R. Eden and B. J. Tatarchuk (2005). Process Integration and Optimization of Logistical Fuels Processing for Hydrogen Production. Computer Aided Chemical Engineering 20B: 1609-1614.

Eljack, F. T. and M. R. Eden (2008). "A systematic visual approach to molecular design via property clusters and group contribution methods." Computers & Chemical Engineering 32(12): 3002-3010.

Eljack, F. T., M. R. Eden, V. Kazantzi, X. Qin and M. M. El-Halwagi (2007). "Simultaneous process and molecular design-A property based approach." AIChE Journal 53(5): 1232-1239.

Elliott, D. J. (2010) Multiscale Modelling. UK Engineering and Physical Sciences Research Council Volume, DOI:

Eriksson, L., E. Johansson and C. Wikström (1998). "Mixture design--design generation, PLS analysis, and model usage." Chemometrics and Intelligent Laboratory Systems 43(1-2): 1.

Ertl, P. (2010). BioPath.Explore, Novartis and Molecular Networks: Biochemical Pathways Database under C@ROL.

Eslick, J. C., Q. Ye, J. Park, E. M. Topp, P. Spencer and K. V. Camarda (2009). "A computational molecular design framework for crosslinked polymer networks." Computers & Chemical Engineering 33(5): 954.

Estrada, E. (1995). "Edge Adjacency Relationships and a Novel Topological Index Related to Molecular Volume." Journal of Chemical Information and Computer Sciences 35(1): 31-3.

Estrada, E. (1995). "Graph theoretical invariant of Randic revisited." Journal of Chemical Information and Computer Sciences 35(6): 1022-5.

Farhang, A. R. and T. Deeter (1999). "Theoretical characterization of phenylene-based oligomers, polymers, and dendrimers." Synthetic Metals 100(1): 141-162.

Faulon, J.-L. (1996). "Stochastic Generator of Chemical Structure. 2. Using Simulated Annealing To Search the Space of Constitutional Isomers." Journal of Chemical Information and Computer Sciences 36(4): 731-740.

Faulon, J.-L., D. P. Visco, Jr. and R. S. Pophale (2003). "The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies." Journal of Chemical Information and Computer Sciences 43(3): 707-720.

Favre, E., L. Marchal-Heusler and M. Kind (2002). "Chemical Product Engineering: Research and Educational Challenges." Chemical Engineering Research and Design 80(1): 65.

Fedors, R. F. (1982). "A relationship between chemical structure and the critical temperature." Chemical Engineering Communications 16(1-6): 149-51.

Fermeglia, M. and S. Pricl (2009). "Multiscale molecular modeling in nanostructured material design and process system engineering." *Computers & Chemical Engineering* 33(10): 1701-1710.

Fransson, M., J. Johansson, A. Sparén and O. Svensson "Comparison of multivariate methods for quantitative determination with transmission Raman spectroscopy in pharmaceutical formulations." *Journal of Chemometrics* 24(12): 674-680.

Fredenslund, A., R. L. Jones and J. M. Prausnitz (1975). "Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures." *AIChE Journal* 21(6): 1086-1099.

Gabrielsson, J., N.-O. Lindberg and T. Lundstedt (2002). "Multivariate methods in pharmaceutical applications." *Journal of Chemometrics* 16(3): 141-160.

Gabrielsson, J., N.-O. Lindberg and T. Lundstedt (2006). "Multivariate Methods in the Development of a New Tablet Formulation: Excipient Mixtures and Principal Properties." *Drug Development and Industrial Pharmacy* 32: 7-20.

Gabrielsson, J., N.-O. Lindberg, M. Palsson, F. Nicklasson, M. Sjostrom and T. Lundstedt (2003). "Multivariate Methods in the Development of a New Tablet Formulation." *Drug Development and Industrial Pharmacy* 29(10): 1053-1075.

Gabrielsson, J., N.-O. Lindberg, M. Palsson, F. Nicklasson, M. Sjostrom and T. Lundstedt (2004). "Optimization and Validation." *Drug Development and Industrial Pharmacy* 30(10): 1037-1049.

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

Gani, R. (2004). "Computer-aided methods and tools for chemical product design." *Chemical Engineering Research and Design* 82(A11): 1494-1504.

Gani, R., L. E. K. Achenie and V. Venkatasubramanian (2003). *Introduction to CAMD. Computer Aided Molecular Design: Theory and Practice*. L. E. K. Achenie, R. Gani and V. Venkatasubramanian. Amsterdam, Elsevier. 12: 3-21.

Gani, R. and E. A. Brignole (1983). "Molecular Design of Solvents for Liquid Extraxction Based on UNIFAC." *Fluid Phase Equilibria* 13: 331-340.

Gani, R. and L. Constantinou (1996). "Molecular structure based estimation of properties for process design." *Fluid Phase Equilibria* 116(1-2): 75-86.

Gani, R., P. M. Harper and M. Hostrup (2005). "Automatic Creation of Missing Groups through Connectivity Index for Pure-Component Property Prediction." *Industrial & Engineering Chemistry Research* 44(18): 7262-7269.

Gani, R., P. M. Harper and M. Hostrup (2010). *ICAS 13, Computer Aided Process Engineering Center*.

Gani, R., G. Hytoft, C. Jaksland and A. K. Jensen (1997). "An integrated computer aided system for integrated design of chemical processes." *Computers & Chemical Engineering* 21(10): 1135-1146.

Gani, R., B. Nielsen and A. Fredenslund (1991). "A group contribution approach to computer-aided molecular design." *AIChE Journal* 37(9): 1318-1332.

Gani, R. and J. P. O'Connell (2001). "Properties and CAPE: from present uses to future challenges." *Computers & Chemical Engineering* 25(1): 3-14.

Gani, R. and E. N. Pistikopoulos (2002). "Property modelling and simulation for product and process design." *Fluid Phase Equilibria* 194-197: 43-59.

Garcia-Munoz S., P. Goulding, A. Balasundaram, J. Mercado, I. Cotto, V. Ruiz, I. Colon-Rivera, D. Rivkees, and D. Sanchez (2010). "Optimal Material Selection of Oral Dosage Development and Manufacture Using Optimization Techniques and Multivariate Latent Variable Models." 2010 AIChE Annual Meeting.

García-Serna, J., L. Pérez-Barrigón and M. J. Cocero (2007). "New trends for design towards sustainability in chemical engineering: Green engineering." *Chemical Engineering Journal* 133(1-3): 7.

Geladi, P. and B. R. Kowalski (1986). "Partial least-squares regression: a tutorial." *Analytica Chimica Acta* 185: 1.

Ghosh, P., A. Sundaram, V. Venkatasubramanian and J. M. Caruthers (2000). "Integrated product engineering: a hybrid evolutionary framework." *Computers & Chemical Engineering* 24(2-7): 685.

Gluck, D. J. (1965). "A Chemical Structure Storage and Search System Developed at Du Pont." *Journal of Chemical Documentation* 5(1): 43.

Gockenbach, M. S. and W. W. Symes (2003). *Adaptive Simulation, the Adjoint State Method, and Optimization, in PDE-Constrained Optimization*, Springer-Verlag.

Gordeeva, E. V., M. S. Molchanova and N. S. Zefirov (1990). "General methodology and computer program for the exhaustive restoring of chemical structures by molecular connectivity indexes. Solution of the inverse problem in QSAR/QSPR." *Tetrahedron Computer Methodology* 3(6B): 389-415.

Groh, W. (1988). "Overtone absorption in macromolecules for polymer optical fibers." *Die Makromolekulare Chemie* 189(12): 2861.

Grooms, D., V. Kazantzi and M. El-Halwagi (2005). "Optimal synthesis and scheduling of hybrid dynamic/steady-state property integration networks." *Computers & Chemical Engineering* 29(11-12): 2318-2325.

Grossmann, I. E. and A. W. Westerberg (2000). "Research challenges in process systems engineering." *AIChE Journal* 46(9): 1700-1703.

Gruber, T. R. (1993). "A translation approach to portable ontology specification." *Knowledge Acquisition* 5(2): 190-220.

Gurdal, Z., R. T. Haftka and P. Hajela (1999). *Design and Optimization of Laminated Composite Materials*. New York, NY, John Wiley & Sons.

Gutman, I. and O. E. Polansky (1986). *Mathematical Concepts in Organic Chemistry*. Berlin, Springer-verlag.

Hall, L. H., R. S. Dailey and L. B. Kier (1993). "Design of molecules from quantitative structure-activity relationship models. 3. Role of higher order path counts: Path 3." *Journal of Chemical Information and Computer Sciences* 33(4): 598.

Hall, L. H., L. B. Kier and J. W. Frazer (1993). "Design of molecules from quantitative structure-activity relationship models. 2. Derivation and proof of information transfer relating equations." *Journal of Chemical Information and Computer Sciences* 33(1): 148.

Hardy, J. and P. McKeown (2007). Ch. 8 - Green Product Design.

Harper, P. M. (2000). *A Multi-Phase, Multi-Level Framework for Computer Aided Molecular Design*. CAPEC, Department of Chemical Engineering, Technical University of Denmark.

Harper, P. M. and R. Gani (2000). "A multi-step and multi-level approach for computer aided molecular design." *Computers & Chemical Engineering* 24(2-7): 677-683.

Harper, P. M., R. Gani, P. Kolar and T. Ishikawa (1999). "Computer-aided molecular design with combined molecular modeling and group contribution." *Fluid Phase Equilibria* 158-160: 337-347.

Heaven, M. C. (1992). "Spectroscopy and Dynamics of Open-Shell van der Waals Molecules." *Annual Review of Physical Chemistry* 43: 283.

Hill, M. (2004). "Product and process design for structured products." *AIChE Journal* 50(8): 1656-1661.

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

Hill, P. J. and K. M. Ng (1997). "Simulation of solids processes accounting for particle-size distribution." *AIChE Journal* 43(3): 715-726.

Horvath, A. L. (1992). *Molecular Design. Chemical Structure Generation from the Properties of Pure Organic Compounds*. *Studies In Physical And Theoretical Chemistry* 1992, 75.

Hosoya, H. (1971). "Topological index. Newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons." *Bulletin of the Chemical Society of Japan* 44(9): 2332-9.

Hostrup, M., R. Gani, Z. Kravanja, A. Sorsak and I. Grossmann (2001). "Integration of thermodynamic insights and MINLP optimization for the synthesis, design and analysis of process flowsheets." *Computers & Chemical Engineering* 25(1): 73-83.

Hostrup, M., P. M. Harper and R. Gani (1999). "Design of environmentally benign processes: integration of solvent design and separation process synthesis." *Computers & Chemical Engineering* 23(10): 1395-1414.

Hutson, J. M. (1990). "Intermolecular Forces from the Spectroscopy of van der Waals Molecules." *Annual Review of Physical Chemistry* 41: 123.

Iachello, F. and R. D. Levine (1995). *Algebraic Theory of Molecules*. New York, NY, Oxford University Press.

Izutsu, K.-i., Y. Hiyama, C. Yomota and T. Kawamishi (2009). "Near-Infrared Analysis of Hydrogen-Bonding in Glass- and Rubber-State Amorphous Saccharide Solids." *AAPS PharmaSciTech* 10(2): 524-529.

Jakobsland, C. and R. Gani (1996). "An integrated approach to process/product design and synthesis based on properties-process relationship." *Computers & Chemical Engineering* 20 (Supplement 1): S151-S156.

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

Joback, K. G. and R. C. Reid (1987). "Estimation of pure-component properties from group contributions." *Chemical Engineering Communications* 57(1-6): 233-43.

Joback, K. G. and G. Stephanopoulos (1989). *Designing molecules possessing desired physical property values*. Proceedings FOCAPD 89, Austin, TX, CACHE corporation.

Joback, K. G. and G. Stephanopoulos (1995). *Searching in Spaces of Discrete Solutions: The Design of Molecules Possessing Desired Physical Properties*. *Advances in Chemical Engineering*, 21, Academic Press. 21.

Johnson, R. A. and D. W. Wichern (2007). *Applied Multivariate Statistical Analysis*. Upper Saddle River, New Jersey, Pearson Prentice Hall.

Jorgensen, S. B., R. Gani, G. Jonsson and M. R. Eden (1999). "Alternative energy-efficient separation processes." *Dansk Kemi* 80(11): 44-47.

Karamertzanis, P. G., A. V. Kazantsev, N. Issa, G. W. A. Welch, C. S. Adjiman, C. C. Pantelides and S. L. Price (2009). "Can the Formation of Pharmaceutical Cocrystals Be Computationally

Predicted? 2. Crystal Structure Prediction." *Journal of Chemical Theory and Computation* 5(5): 1432.

Karamertzanis, P. G. and S. L. Price (2006). "Energy Minimization of Crystal Structures Containing Flexible Molecules." *Journal of Chemical Theory and Computation* 2(4): 1184.

Karunanithi, A. T., L. E. K. Achenie and R. Gani (2005). "A New Decomposition-Based Computer-Aided Molecular/Mixture Design Methodology for the Design of Optimal Solvents and Solvent Mixtures." *Industrial & Engineering Chemistry Research* 44(13): 4785-4797.

Karunanithi, A. T., C. Acquah and L. E. K. Achenie (2008). "Tuning the Morphology of Pharmaceutical Compounds via Model Based Solvent Selection." *Chinese Journal of Chemical Engineering* 16(3): 465.

Karunanithi, A. T., C. Acquah, L. E. K. Achenie, S. Sithambaram, S. L. Suib and R. Gani (2007). "An experimental verification of morphology of ibuprofen crystals from CAMD designed solvent." *Chemical Engineering Science* 62(12): 3276-3281.

Katritzky, A. R. and E. V. Gordeeva (1993). "Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research." *Journal of Chemical Information and Computer Sciences* 33(6): 835-57.

Kazantzi, V., X. Qin, M. El-Halwagi, F. Eljack and M. Eden (2007). "Simultaneous Process and Molecular Design through Property Clustering Techniques: A Visualization Tool." *Industrial & Engineering Chemistry Research* 46(10): 3400-3409.

Kehiaian, H. V. (1983). "Group contribution methods for liquid mixtures: a critical review." *Fluid Phase Equilibria* 13: 243-52.

Kepner, C. H. and B. B. Tregoe (1997). *The New Rational Manager: An Updated Edition for a New World*. Princeton, New Jersey, Princeton Research Press.

Kettaneh-Wold, N. (1991). "Use of experimental design in the pharmaceutical industry." *Journal of Pharmaceutical and Biomedical Analysis* 9(8): 605.

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

Kier, L. B. (1985). "A shape index from molecular graphs." *Quantitative Structure-Activity Relationships* 4(3): 109-16.

Kier, L. B. and L. H. Hall (1976). "Molecular connectivity VII. Specific treatment of heteroatoms." *Journal of pharmaceutical sciences* 65(12): 1806-9.

Kier, L. B. and L. H. Hall (1986). *Chemometrics Series, 9: Molecular Connectivity in Structure-Activity Analysis*. New York, John Wiley & Sons.

Kier, L. B., L. H. Hall and J. W. Frazer (1993). "Design of molecules from quantitative structure-activity relationship models. 1. Information transfer between path and vertex degree counts." *Journal of Chemical Information and Computer Sciences* 33(1): 143-7.

Klein, C. (2002). *Mineral Science*, John Wiley & Sons.

Klein, J. A., D. T. Wu and R. Gani (1992). "Computer aided mixture design with specified property constraints." *Computers & Chemical Engineering* 16(5): 229-236.

Koch, R. (1982). "Molecular connectivity and acute toxicity of environmental pollutants." *Chemosphere FIELD Full Journal Title:Chemosphere* 11(9): 925-31.

Kokossis, A. C. and Y. Yang (2009). "Future system challenges in the design of renewable bio-energy systems and the systems of sustainable biorefineries." *Design for Energy and Environment*, 107-123.

Kondo, T. (1997). "The Relationship between Intramolecular Hydrogen Bonds and Certain Physical Properties of Regioselectively Substituted Cellulose Derivatives." *Journal of Polymer Science Part B: Polymer Physics* 35(4): 717-723.

Kramer, R. (1998). *Chemometric Techniques for Quantitative Analysis*. New York, NY, Marcel Dekker.

Kvasnicka, V. and J. Pospichal (1990). "Canonical indexing and constructive enumeration of molecular graphs." *Journal of Chemical Information and Computer Sciences* 30(2): 99-105.

Kvasnicka, V. and J. Pospichal (1996). "Simulated Annealing Construction of Molecular Graphs with Required Properties." *Journal of Chemical Information and Computer Sciences* 36(3): 516-26.

Land, A. H. and A. G. Doig (1960). "An Automatic Method of Solving Discrete Programming Problems." *Econometrica* 28(3): 497.

Larsen, P. V. (2003). "ST111: Regression analysis and analysis of variance." from <http://statmaster.sdu.dk/courses/st111/>.

Le Riche, R. and R. T. Haftka (1995). "Improved genetic algorithm for minimum thickness composite laminate design." *Composite Engineering* 5(2): 143-161.

Leach, A. R. (2001). *Molecular Modelling: Principles and Applications*. New York, NY, Prentice Hall.

Liao, J. C., R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti and V. P. Roychowdhury (2003). "Network component analysis: Reconstruction of regulatory signals in biological systems." *Proceedings of the National Academy of Sciences of the United States of America* 100(26): 15522-15527.

Maghsoodloo, S. and J. N. Hool (1984). "On the Relationship of General Designs and Moments of Estimators in Simple Linear Regression." *IIE Transactions* 16(2).

Marcoulaki, E. C. and A. C. Kokossis (1998). "Molecular design synthesis using stochastic optimisation as a tool for scoping and screening." *Computers & Chemical Engineering* 22 (Supplement 1): S11-S18.

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

Martin, T. M. and D. M. Young (2001). "Prediction of the Acute Toxicity (96-h LC50) of Organic Compounds to the Fathead Minnow (*Pimephales promelas*) Using a Group Contribution Method." *Chemical Research in Toxicology* 14(10): 1378-1385.

Martinello, T., T. M. Kaneko, M. V. R. Velasco, M. E. S. Taqueda and V. O. Consiglieri (2006). "Optimization of poorly compactable drug tablets manufactured by direct compression using the mixture experimental design." *International Journal of Pharmaceutics* 322(1-2): 87.

Mavrovouniotis, M. L. (1990). "Estimation of properties from conjugate forms of molecular structures: the ABC approach." *Industrial & Engineering Chemistry Research* 29(9): 1943-53.

Mayo, S. L., B. D. Olafson and W. A. Goddard, III (1990). "DREIDING: a generic force field for molecular simulations." *Journal of Physical Chemistry* 94(26): 8897-909.

McAllister, S. R. and C. A. Floudas (2008). "[alpha]-Helical Topology Prediction and Generation of Distance Restraints in Membrane Proteins." *Biophysical Journal* 95(11): 5281.

McKenzie, A. (2002). "The Handbook for Organic Chemistry Lab." 4th. Retrieved Sept. 16, 2010, from <http://orgchem.colorado.edu/hndbksupport/irtutor/IRtheory.pdf>.

McMurry, J. (1995). *Organic Chemistry*, Brooks & Cole Publishing.

Meyer, M. H. and D. Dalal (2003). "Managing platform architectures and manufacturing processes for nonassembled products." *Journal of Product Innovation Management* 19: 277-293.

Meyer, R. K. and C. J. Nachtsheim (1995). "The Coordinate-Exchange Algorithm for Constructing Exact Optimal Experimental Designs." *Technometrics* 37(1): 60-69.

Mirau, P. A. (1993). *NMR Characterization of Polymers*. Polymer Characterization. B. J. Hunt and M. I. James, Chapman & Hall.

Montgomery, D. C. (2005). *Design and Analysis of Experiments*. USA, John Wiley & Sons.

Montgomery, D. C. and R. H. Meyers (1995). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. New York, NY, John Wiley & Sons.

Murray, I. (1987). *The NIR Spectra of Homologous Series of Organic Compounds in Near-Infrared Diffuse Reflectance/Transmittance Spectroscopy*. Budapest.

Murray, W. J., L. H. Hall and L. B. Kier (1975). "Molecular connectivity. III: Relationship to partition coefficients." *Journal of pharmaceutical sciences* 64(12): 1978-81.

Muteki, K. and J. F. MacGregor (2007). "Multi-Block PLS modeling for L-shape data structures with applications to mixture modeling." *Chemometrics and Intelligent Laboratory Systems* 85: 186-194.

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

Muteki, K., J. F. MacGregor and T. Ueda (2007). "Mixture designs and models for the simultaneous selection of ingredients and their ratios." *Chemometrics and Intelligent Laboratory Systems* 86: 17-25.

Nadler, T. K., S. T. McDaniel, M. O. Westerhaus and J. S. Shenk (1989). "Deconvolution of Near-Infrared Spectra." *Applied Spectroscopy Reviews* 43(8): 1354-1358.

Naes, T. and T. Isaksson (1992). *Computer-Assisted Methods in Near-Infrared Spectroscopy. Computer-Enhanced Analytical Spectroscopy*. P. Jurs. New York, NY, Plenum Press. 3.

Ndindayino, F., D. Henrist, F. Kiekens, C. Vervact and J. P. Remon (1999). "Characterization and evaluation of isomalt performance in direct compression." *International Journal of Pharmaceutics*(189): 113-124.

Nesbitt, D. J. (1988). "High Resolution Infrared Spectroscopy of Weakly Bound Molecular Complexes." *Chemical Reviews* 88: 843.

Ng, K. M., R. Gani and W. D. Seider (2009). "Special issue on "Chemical products: from conceptualization to commercialization"." *Computers & Chemical Engineering* 33(5): 929.

Nielsen, T. L. and R. Gani (2001). "A computer aided system for correlation and prediction of phase equilibrium data." *Fluid Phase Equilibria* 185(1-2): 13-20.

Nikonenko, N. A., D. K. Buslov, N. I. Sushko and R. G. Zhbankov (2002). "Analysis of the structure of carbohydrates with use of the regularized deconvolution method of vibrational spectra." *BAU Fen Bil. Enst. Dergisi* 4(2): 13-16.

Norris, T., P. K. Aldridge and S. S. Sekulic (1997). "Determination of End-points for Polymorph Conversions of Crystalline Organic Compounds Using On-line Near-Infrared Spectroscopy." *Analyst* 122: 549-552.

Noureldin, M. B. and M. M. El-Halwagi (2000). "Pollution prevention targets through integrated design and operation." *Computers & Chemical Engineering* 24(2-7): 1445-1453.

Noy, N. F. and D. L. McGuinness (2001). "Ontologoy Development 101: A guide to creating your first ontology." *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05*.

O'Connell, J. P., R. Gani, P. M. Mathias, G. Maurer, J. D. Olson and P. A. Crafts (2009). "Thermodynamic Property Modeling for Chemical Process and Product Engineering: Some Perspectives." *Industrial & Engineering Chemistry Research* 48(10): 4619.

Odjo, A. O., N. E. Sammons Jr., W. Yuan, A. Marcilla, M. R. Eden, J. A. Caballero. (2011). "A Disjunctive-Genetic Programming Approach to Synthesis of Process Networks", *Industrial & Engineering Chemistry Research* 50(10): 6213-6228.

O'Neil, A. J., R. D. Jee and A. C. Moffat (1998). "The application of multiple linear regression to the measurement of the median particle size of drugs and pharmaceutical excipients by near-infrared spectroscopy." *Analyst* 123: 2297-2302.

Olson, G. B. (1997). "Computational Design of Hierarchically Structured Materials." *Science* 277(5330): 1237-1242.

Ortiga, J., J. C. Bruno, A. Coronas, I. E. Grossman and A. Valentin Plesu and Paul Serban (2007). Review of optimization models for the design of polygeneration systems in district heating and cooling networks. *Computer Aided Chemical Engineering*, Elsevier. Volume 24: 1121.

Ortiz, P. J. and C. S. Perez (1982). *Elementos de mecanica cuantica y estructura atomica*, Universidad de La Babana.

Ostrovsky, G. M., L. E. K. Achenie and M. Sinha (2002). "On the solution of mixed-integer nonlinear programming models for computer aided molecular design." *Computers & Chemistry* 26(6): 645-660.

Parker, J. S. and J. D. Moseley (2008). "Kepner-Tregoe Decision Analysis as a Tool to Aid Route Selection. Part 1." *Organic Process Research & Development* 12(6): 1041.

Parthasarathy, G. and M. M. El-Halwagi (2000). "Optimum mass integration strategies for condensation and allocation of multicomponent VOCs." *Chemical Engineering Science* 55(5): 881-895.

Pretel, E. J., P. A. Lopez, S. B. Bottini, E. A. Brignole (1994). "Computer-aided molecular design of solvents for separation processes." *AIChE Journal* 40(8): 1349-1360.

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): 3792-3798.

Radhakrishnan, V. K., E. W. Davis and V. A. Davis (2010). "Influence of initial mixing methods on melt-extruded single-walled carbon nanotube-polypropylene nanocomposites." *Polymer Engineering & Science* 50(9): 1831-1842.

Raman, V. S. and C. D. Maranas (1998). "Optimization in product design with properties correlated with topological indices." *Computers & Chemical Engineering* 22(6): 747.

Randic, M. (1975). "Characterization of molecular branching." *Journal of the American Chemical Society* 97(23): 6609-15.

Randic, M. and S. C. Basak (2001). "A new descriptor for structure-property and structure-activity correlations." *Journal of Chemical Information and Computer Sciences* 41(3): 650-6.

Randic, M. and J. Zupan (2001). "On Interpretation of Well-Known Topological Indices." *Journal of Chemical Information and Computer Sciences* 41(3): 550-560.

Ray, L. C. and R. A. Kirsch (1957). "Finding Chemical Records by Digital Computers." *Science*(126): 814-819.

Reid, R. C., J. M. Prausnitz and B. E. Pauling (1987). *The properties of gases and liquids*. New York, McGraw-Hill.

Rodriguez, F., C. Cohen, C. K. Ober and L. A. Archer (2003). *Principles of Polymer Systems*. New York, NY, Taylor & Francis.

Russel, B. M., J. P. Henriksen, S. B. Jørgensen and R. Gani (2000). "Integration of design and control through model analysis." *Computers & Chemical Engineering* 24(2-7): 967.

Sahinidis, N. V., M. Tawarmalani and M. Yu (2003). "Design of alternative refrigerants via global optimization." *AIChE Journal* 49(7): 1761-1775.

Salari, A. and R. E. Young (1998). "Application of attenuated total reflectance FTIR spectroscopy to the analysis of mixtures of pharmaceutical polymorphs." *International Journal of Pharmaceutics*(163): 157-166.

Sammons Jr., N. E. (2009). *A Framework for Optimal Biomass-based Polygeneration Facility Product Allocation*. Ph.D. Dissertation, Dept. of Chemical Engineering. Auburn University

Satyanarayana, K. C., J. Abildskov and R. Gani (2007). "Simultaneous polymer property modeling using grid technology for structured products." *Computer-Aided Chemical Engineering* 24: 321-326.

Satyanarayana, K. C., J. Abildskov and R. Gani (2009). "Computer-aided polymer design using group contribution plus property models." *Computers & Chemical Engineering* 33(5): 1004-1013.

Satyanarayana, K. C., R. Gani and J. Abildskov (2007). "Polymer property modeling using grid technology for design of structured products." *Fluid Phase Equilibria* 261(1-2): 58-63.

Saykally, R. J. (1989). "Far-Infrared Laser Spectroscopy of van der Waals Bonds: A Powerful New Probe of Intermolecular Forces." *Accounts of Chemical Research* 22: 295.

Scheffe, H. (1958). "Experiments with mixtures." *J. R. Stat. Soc. B.* 20(2): 344-360.

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

Schmid, E. F. and D. A. Smith (2002). "Should scientific innovation be managed?" Drug Discovery Today 7(18): 941.

Schmit, L. A. and B. Farshi (1973). "Optimum laminate design for strength and stiffness." International Journal for Numerical Methods in Engineering 7(4): 519-536.

Seider, W. D., S. Widagdo, J. D. Seader and D. R. Lewin (2009). "Perspectives on chemical product and process design." Computers & Chemical Engineering 33(5): 930.

Seitz, J. T. (1993). "The estimation of mechanical properties of polymers from molecular structure." Journal of Applied Polymer Science 49(8): 1331-1351.

Seyer, J. J., P. E. Luner and M. S. Kemper (2000). "Application of Diffuse Reflectance Near-Infrared Spectroscopy for Determination of Crystallinity." Journal of Pharmaceutical Sciences 89(10): 1305-1316.

Sharma, C. V. K. (2002). "Crystal Engineering - Where Do We Go from Here?" Crystal Growth & Design 2(6): 465-474.

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

Sheridan, R. P. and S. K. Kearsley (1995). "Using a Genetic Algorithm To Suggest Combinatorial Libraries." Journal of Chemical Information and Computer Sciences 35(2): 310-20.

Sinha, M. and L. E. K. Achenie (2001). "Systematic design of blanket wash solvents with recovery considerations." Advances in Environmental Research 5(3): 239-249.

Sinha, M., L. E. K. Achenie and R. Gani (2003). "Blanket wash solvent blend design using interval analysis." Industrial & Engineering Chemistry Research 42(3): 516-527.

Sinha, M., L. E. K. Achenie and G. M. Ostrovsky (1999). "Environmentally benign solvent design by global optimization." Computers & Chemical Engineering 23(10): 1381-1394.

Siqueira Petri, D. F., G. Wenz, P. Schunk and T. Schimmel (1999). "An Improved Method for the Assembly of Amino-Terminated Monolayers on SiO₂ and the Vapor Deposition of Gold Layers." Langmuir 15(13): 4520.

Skvortsova, M. I., I. I. Baskin, O. L. Slovokhotova, V. A. Palyulin and N. S. Zefirov (1993). "Inverse problem in QSAR/QSPR studies for the case of topological indexes characterizing molecular shape (Kier indices)." Journal of Chemical Information and Computer Sciences 33(4): 630-4.

Smith, B. V. and M. Ierapetritou (2009). "Framework for Consumer-Integrated Optimal Product Design." *Industrial & Engineering Chemistry Research* 48(18): 8566.

Smith, J. M., H. C. V. Ness and M. M. Abbott (1996). *Introduction to Chemical Engineering Thermodynamics*. New York, NY, McGraw-Hill Companies, Inc.

Smith, W. F. and T. A. Beverly (1997). "Generating linear and quadratic Cox mixture models." *Journal of Quality Technology* 29: 211-224.

Socrates, G. (2001). *Infrared and Raman Characteristic Group Frequencies*. New York, NY, John Wiley & Sons, Ltd.

Solvason, C. C., N. G. Chemmangattuvalappil and M. R. Eden (2010). "Multi-Scale Chemical Product Design using the Reverse Problem Formulation." *Computer Aided Chemical Engineering* 28: 1285-1290.

Solvason, C. C., N. G. Chemmangattuvalappil and M. R. Eden (2009). "Decomposition techniques for multiscale structured product design: molecular synthesis." *Computer-Aided Chemical Engineering* 26: 153-158.

Solvason, C. C., N. G. Chemmangattuvalappil and M. R. Eden (2009). "A systematic method for integrating product attributes within molecular synthesis." *Computers & Chemical Engineering* 33(5): 977-991.

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

Solvason, C. C., F. T. Eljack, N. G. Chemmangattuvalappil and M. R. Eden (2008). "Visual Mixture Design Using Property Clustering." *Computer Aided Chemical Engineering* 25: Paper 442 (CD Volume).

Batsy D. R., C. C. Solvason, N. E. Sammons Jr., V. Chambost, M. R. Eden, M. M. El-Halwagi and P. R. Stuart (2010). "Techniques for Product Portfolio Selection and Process Design of the Forest Biorefinery" chapter in *Integrated Biorefineries: Design, Analysis, and Optimization*, P. R. Stuart and M. M. El-Halwagi Eds. CRC Press (accepted for publication).

Sorensen, A. H., J. M. Sonnergaard and L. Hovgaard (2006). "Bulk Characterization of Pharmaceutical Powders by Low-Pressure Compression II: Effect of Method Settings and Particle Size." *Pharmaceutical Development and Technology* 11: 235-241.

Srinivas, B. K. and M. M. El-Halwagi (1993). "Optimal design of pervaporation systems for waste reduction." *Computers & Chemical Engineering* 17(10): 957-970.

Stephanopoulos, G. (2003). *Invention and innovation in a Product-Centered Chemical Industry: General trends and a case study*. AIChE, San Francisco, CA.

Stephenson, G. A., R. A. Forbes and S. M. Reutzel-Edens (2001). "Characterization of the solid state: quantitative issues." *Advanced Drug Delivery Reviews* 48: 67-90.

Street, C., J. Woody, J. Ardila and M. Bagajewicz (2008). "Product Design: A Case Study of Slow-Release Carpet Deodorizers/Disinfectants." *Industrial & Engineering Chemistry Research* 47(4): 1192.

Tolle, I., X. Huang, Y. A. Akpalu and L. L. Martin (2009). "A Modified Network Component Analysis (NCA) Methodology for the Decomposition of X-ray Scattering Signatures." *Industrial & Engineering Chemistry Research* 48(13): 6137.

Tosi, C. and A. Pinto (1972). "Near-infrared spectroscopy of hydrocarbon functional groups." *Spectrochim. Acta*. 28A: 585-597.

Trinajstić, N. (1992). *Chemical Graph Theory*. Boca Raton, FL, CRC press.

Tsolou, G., V. G. Mavrantzas, Z. A. Makrodimitri, I. G. Economou and R. Gani (2008). "Atomistic Simulation of the Sorption of Small Gas Molecules in Polyisobutylene." *Macromolecules* 41(16): 6228.

Uchida, M. (1980). "Affinity and mobility of fungicidal dialkyl dithiolanylidene malonates in rice plants." *Pesticide Biochemistry and Physiology* 14(3): 249.

Ugi, I., J. Bauer, K. Bley, A. Dengler, A. Dietz, E. Fontain, B. Gruber, R. Herges, M. Knauer, K. Reitsam and N. Stein (1993). "Computer-Assisted Solution of Chemical Problems - The Historical Development and the Present State of the Art of a New Discipline of Chemistry." *Angewandte Chemie International Edition in English* 32(2): 201-227.

Ugi, I., J. Bauer, C. Blomberger, J. Brandt, A. Dietz, E. Fontain, B. Gruber, A. v. Scholley-Pfab, A. Senff and N. Stein (1994). "Models, concepts, theories, and formal languages in chemistry and their use as a basis for computer assistance in chemistry." *Journal of Chemical Information and Computer Sciences* 34(1): 3.

Ugi, I., J. Bauer, J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum and W. Schubert (1979). "New Applications of Computers in Chemistry." *Angewandte Chemie International Edition in English* 18(2): 111-123.

Uhlemann, J. and I. Reiß "Product Design and Process Engineering using the Example of Flavors." *Chemical Engineering & Technology* 33(2): 199-212.

Ulas, S. and U. M. Diwekar (2006). "Integrating product and process design with optimal control: A case study of solvent recycling." *Chemical Engineering Science* 61(6): 2001.

Utterback, J. (1994). *Mastering the dynamics of innovation*, Harvard Business School Press.

Vaidyanathan, R. and M. El-Halwagi (1994). "Global optimization of nonconvex nonlinear programs via interval analysis." *Computers & Chemical Engineering* 18(10): 889-897.

Vaidyanathan, R., Y. Gawayed and M. El-Halwagi (1998). "Computer-aided design of fiber reinforced polymer composite products." *Computers & Chemical Engineering* 22(6): 801-808.

Van Krevelen, D. W. (1990). *Properties of polymers*. Amsterdam, Elsevier.

Vandenberg, E. T., L. Bertilsson, B. Liedberg, K. Uvdal, R. Erlandsson, H. Elwing and I. Lundström (1991). "Structure of 3-aminopropyl triethoxy silane on silicon oxide." *Journal of Colloid and Interface Science* 147(1): 103.

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

Venkatasubramanian, V., K. Chan and J. M. Caruthers (1994). "Computer-aided molecular design using genetic algorithms." *Computers & Chemical Engineering* 18(9): 833-44.

Venkatasubramanian, V., K. Chan and J. M. Caruthers (1995). "Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm." *Journal of Chemical Information and Computer Sciences* 35(2): 188-95.

Venkatasubramanian, V., C. Zhao, G. Joglekar, A. Jain, L. Hailemariam, P. Suresh, P. Akkisetty, K. Morris and G. V. Reklaitis (2006). "Ontological informatics infrastructure for pharmaceutical product development and manufacturing." *Computers & Chemical Engineering* 30(10-12): 1482.

Vladutz, G. E. (1963). "Concerning one System of Classification and Codification of Organic Reactions." *Information Storage and Retrieval* 1: 117-146.

Wachter, A. and L. T. Biegler (2006). "on the implementation of interior-point filter line-search algorithm for large-scale nonlinear programming." *Mathematical Programming*(106): 25-27.

Wang, R., Y. Fu and L. Lai (1997). "A New Atom-Additive Method for Calculating Partition Coefficients." *Journal of Chemical Information and Computer Sciences* 37(3): 615.

Wang, Y. and L. E. K. Achenie (2002). "A hybrid global optimization approach for solvent design." *Computers & Chemical Engineering* 26(10): 1415-1425.

Wei, J. (1996). "A century of changing paradigms in chemical engineering." *ChemTech* 26(5): 16.

Weis, D. C., J.-L. Faulon, R. C. LeBorne and D. P. Visco, Jr. (2005). "The Signature Molecular Descriptor. 5. The Design of Hydrofluoroether Foam Blowing Agents Using Inverse-QSAR." *Industrial & Engineering Chemistry Research* 44(23): 8883-8891.

Werpy, T. and G. Petersen (2004). *Top Value Added Chemicals From Biomass*. D. o. Energy, Pacific Northwest National Laboratory and National Renewable Energy Laboratory.

Westerberg, A. W. (2004). "A retrospective on design and process synthesis." *Computers & Chemical Engineering* 28(4): 447.

Westerberg, A. W. and E. Subrahmanian (2000). "Product design." *Computers & Chemical Engineering* 24(2-7): 959.

Wibowo, C. and K. M. Ng (2001). "Product-oriented process synthesis and development: Creams and pastes." *AIChE Journal* 47(12): 2746-2767.

Wibowo, C. and K. M. Ng (2002). "Product-centered processing: Manufacture of chemical-based consumer products." *AIChE Journal* 48(6): 1212-1230.

Wiener, H. (1947). "Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons." *Journal of the American Chemical Society* 69: 2636-8.

Williams, A. J. (2007). ChemSpider, Royal Society of Chemistry.

Wilson, R. J. (1986). *Introduction to Graph Theory*, Longman Scientific & Technical.

Wold, S., K. Esbensen and P. Geladi (1987). "Principal Component Analysis." *Chemometrics and Intelligent Laboratory Systems* 2: 37-52.

Wold, S., P. Geladi, K. Esbensen and J. Öhman (1987). "Multi-way principal components-and PLS-analysis." *Journal of Chemometrics* 1(1): 41-56.

Wong, J. S. and C. B. Moore (1982). "Inequivalent C-H oscillators of gaseous alkanes and alkenes in laser photoacoustic overtone spectroscopy." *Journal of Physical Chemistry* 77(2): 602-615.

Workman, J. and A. Springsteen (1998). *Applied Spectroscopy: A compact Reference for Practitioners*. San Diego, CA, Academic Press.

Workman, J. and L. Weyer (2008). *Practical Guide to Interpretive Near-Infrared Spectroscopy*. New York, NY, CRC Press.

Wu, H. S. and S. I. Sandler (1989). "Proximity effects on the predictions of the UNIFAC model: I. Ethers." *AIChE Journal* 35(1): 168-72.

Wu, H. S. and S. I. Sandler (1991). "Use of ab initio quantum mechanics calculations in group contribution methods. 1. Theory and the basis for group identifications." *Industrial & Engineering Chemistry Research* 30(5): 881-9.

Wu, T., X. Zhou and S.-H. Yu (2009). *The Research Group Decision Methods for Product Design Evaluation*. Fourth International Conference on Computer Sciences and Convergence Information Technology: Applied Ergonomics, Seoul, Korea.

Xin, Y. and W. B. Whiting (2000). "Case Studies of Computer-Aided Design Sensitivity to Thermodynamic Data and Models." *Industrial & Engineering Chemistry Research* 39(8): 2998.

Young, D., R. Scharp and H. Cabezas (2000). "The waste reduction (WAR) algorithm: environmental impacts, energy consumption, and engineering economics." *Waste Management* 20(8): 605.

Young, D. M. (2001). *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*. New York, NY, John Wiley & Sons.

Yu, L., S. M. Reutzel and G. A. Stephenson (1998). "Physical Characterization of polymorphic drugs: an integrated characterization strategy." *PSST* 1(3): 118-127.

Zemany, P. D. (1950). "Punched Card Catalog of Mass Spectra Useful in Quantitative Analysis." *Analytical Chemistry*(22): 920-922.

Zeng, Q. H., A. B. Yu and G. Q. Lu (2008). "Multiscale modeling and simulation of polymer nanocomposites." *Progress in Polymer Science* 33(2): 191.

Zerbi, G., F. Ciampelli and V. Zamboni (1965). "Classification of Crystallinity Bands in the Infrared Spectra of Polymers." *Journal of Polymer Science: Part C* 7: 141-151.

Zhang, F. and M. P. Srinivasan (2004). "Self-Assembled Molecular Films of Aminosilanes and Their Immobilization Capacities." *Langmuir* 20(6): 2309.

Zhu, M., M. M. El-Halwagi and M. Al-Ahmad (1997). "Optimal design and scheduling of flexible reverse osmosis networks." *Journal of Membrane Science* 129(2): 161-174.

Appendix A1

Additional Chemical Product Design Techniques

This section contains other useful tools for the design of chemical products and processes, including multicriteria decision making to aid in the stage-gate development process, ontology to aid in data gathering and interpretation, and mathematical optimization to aid in efficient solution derivation.

A1.1. Multi-Criteria Decision Making (MCDM)

One of simplest methods for converting data and information into workable products was proposed by Dym and Little (2008). In this method, information is first obtained from the client, who is either the customer or end-use consumer. The information is then transformed to a set of goals or attributes which may cover overviews of form, function, costs, potential markets, and safety issues. Next, these goals are weighted by importance and impact on the product. Those goals deemed most important and carrying the highest impact are kept in the design. Products that have these important attributes are then selected. When this process is systematically applied, it is referred to as decision analysis (Kepner and Tregoe 1997) or multi-criteria decision making (MCDM). Although the idea of using subjective decision analysis is far from perfect, it has several advantages (Cussler and Moggridge 2001)

1. The selection of feasible products is typically a point of fierce management review and will require the support of managers outside of the core product development team. It can become an excellent tool for garnering support for the product.
2. The need to weight and score each idea make the consulting of experts and customers inevitable while numerical scoring trends make it difficult for a single view to dominate decisions.
3. The separate scoring of different criteria ensures that strengths and weaknesses of each product are obvious.

An example MCDM process called Kepner-Tregoe Decision Analysis was used by AstraZeneca to aid in the design of the synthesis route for the pyruvate dehydrogenase kinase (PDK) inhibitor, AZD7545, as shown in Table A1.1. The first step of the analysis was to develop a succinct objective, such as “To identify routes for the long-term manufacture of the required drug candidate, paying particular attention to critical factors (e.g., safety, environmental, cost, intellectual property, etc.)” (Parker and Moseley 2008) Next, a list of product criteria were generated and classified as either “musts” or “wants.” Must criteria were critical issues without which the product could not be made. Want criteria were comprised of all other objectives. The wants are then weighted between 0 and 10, and, for this example, chemical feasibility was given the highest value indicating it was the most important want criteria.

Table A1.1: An Example MCDM for the Route Identification of PDK Inhibitor, AZD7545 (Parker and Moseley 2008).

route no.	route description (name)	Musts					go/no go	Wants						total score	initial rank
		safety	health	environment	chemical feasibility	raw materials		chemical feasibility, 10 ^a	cost, 8 ^a	environment, 8 ^a	accommodation, 6 ^a	numbers of steps, 4 ^a			
Initial Routes															
1.1	medicinal chemistry route	1	1	0	1	0	no	4.25	2.1	0	4	3	95	34	
1.2	initial contractor's route	1	1	0	1	1	no	4.59	2.5	0	4	2	98	31	
Campaign 2 Route Variations															
2.1	contractor's campaign 2 route	1	1	1	1	1	yes	5.00	4.5	4	3	2	144	1	
2.2	with catalytic Pd	1	1	1	1	1	yes	4.40	4.5	4	3	2	139	3	
2.3	with catalytic other metal	1	1	1	1	1	yes	3.00	4.5	4	3	2	124	9	
2.4	nonmetal route	1	1	1	1	1	yes	3.00	4.5	5	3	2	132	6	
2.5	with oxidation first	1	1	1	1	1	yes	3.56	5.0	4	3	2	134	4	
2.6	nitro thiol route	1	1	1	1	1	yes	0.51	4.5	3	4	1	93	35	
Alternative Leaving Group Variations															
3.1	diazonium LG	1	1	1	1	1	yes	1.20	4.5	4	3	1	102	25	
3.2	tosylate LG	1	1	1	1	1	yes	0.86	4.5	5	5	1	119	12	
3.3	nitro LG	1	1	1	1	1	yes	2.93	4.5	5	5	2	143	2	
3.4	nitrile route	1	1	1	1	1	yes	2.05	4.5	5	5	1	130	7	
3.5	nitro/diazo route	?? ^b	1	1	??	1	??	not scored					n/s	40	
Selected Other Route Options															
4.1	quetiapine-type best route	1	1	1	1	0	no	3.42	4.5	4	4	2	134	4	
5.4	tosyl chloride and aniline	1	1	1	1	1	yes	1.07	5.0	3	4	3	111	16	
5.6	typical tosyl chloride based route	1	1	1	1	1	yes	1.37	5.0	3	3	2	104	23	
5.9	sulfobenzoic acid and aniline	1	1	1	1	1	yes	1.68	5.0	4	4	4	129	8	
6.2	tolyl Grignard with protected amide	1	1	1	1	1	yes	1.96	2.0	4	4	3	103	24	
7.2	protected aniline Grignard	1	1	1	1	1	yes	2.13	2.2	4	4	3	107	19	
8.3	sulfinic acid (lowest scored route)	1	1	1	??	1	??	0.36	3.2	2	3	2	70	38	
9.1	best bis-aryl sulfone route	1	1	1	1	1	yes	0.77	5.0	4	5	2	118	13	
^a Weighting. ^b Undecided/unknown.															

^a Weighting. ^b Undecided/unknown.

The next step of the process was to generate alternatives, or potential solutions to the objectives. In the case of the example presented, different synthesis routes were identified by expertise, scientific journal review, or computational methods, using an enzyme classification system. Twenty-one types of routes were evaluated. It should be noted that if the product was more complex (i.e. contained molecular architecture at multiple length scales), it would have had a longer list of alternatives. The first two alternatives failed to meet the must criteria so they were eliminated immediately. The 19 remaining possibilities were evaluated on their want criteria, which was the most difficult to perform because of their highly subjective nature. In order to minimize the subjective nature, it was important to decide on the scoring method prior to determining which alternatives to evaluate.

As shown in Table A1.2 the alternatives were scored against 5 wants: accommodation, chemical feasibility, costs of goods, environment, and number of stages. Additional want criteria could have been thermodynamic and kinetic barriers, reaction yields, etc. A default score of 5 was given to each alternative for each want criteria, with subtractions made for those alternatives with less than desirable characteristics. In some cases, it was necessary to identify adverse consequences resulting from some of the alternatives, such as chemical feasibility and undesired inter-dependencies.

Table A1.2: Scoring System for Wants in the Route Identification of PDK Inhibitor, AZD7545 (Parker and Moseley 2008)

Want	scoring system
accommodation	default score 5 –1 for each stage requiring special accommodation from standard pilot plant vessels (e.g., cryogenic, hydrogenation or pressure vessels)
chemical feasibility	default score normalised to 5 for best route known or estimated yield for each stage multiplied by confidence as assessed by rating below; total for all stages multiplied together and normalised to score of 5 for best route, all others pro rata score 10 for exact literature or AstraZeneca precedent score 8 for close literature precedent score 6 for a “better than even chance” score 4 for significant doubts, but some precedent score 2 for “one obscure paper suggests this is possible”
costs of goods	default score normalised to 5 for best route known or estimated yields for all stages after introduction of chiral acid multiplied together and normalised to score of 5 for best route, all others pro rata
environment	default score 5 –1 for each stage requiring special abatement or disposal (e.g., catalytic heavy metals, chlorination, iodine waste, oxidation)
number of stages	score 5 for 3 steps score 4 for 4 steps score 3 for 5 steps score 2 for 6 steps score 1 for 7 steps score 0 for 8 or more steps

Finally, the criteria were gathered for the decision analysis matrix and were calculated by adding all of the want criteria scores together for each alternative. The alternatives with the highest scores were chosen for further analysis, including experimental product prototyping and algorithmic process development. The lowest scored alternatives are removed from contention along with any criteria that failed the must criteria. In the PDK Inhibitor AZD7545 example, laboratory scale prototyping and product synthesis was performed and the scores were adjusted to reflect the additional source of information. In particular, route 3.3, which substituted a nitro-group for the bromide in brombenzamide, was initially the second-best option identified. However, no coupling reaction between thocyanate and 4-nitrobenzamide, which had a known pathway, could be established, so it then failed the feasibility must criteria, and was removed from consideration. The final, post-feasibility, pre-development list of options are presented in Table A1.3.

Table A1.3: The final MCDM for the Route Identification of PDK Inhibitor, AZD7545 (Parker and Moseley 2008).

route no.	route description (name)	musts		go/no go	initial score	initial rank	final score	change in score	final rank ^a
		chemical feasibility	raw materials						
Initial Routes									
1.1	medicinal chemistry route	1	0	no	95	34	98	+3	n/a
1.2	initial contractor's route	1	1	no	98	31	98	0	n/a
Campaign 2 Route Variations									
2.1	contractor's campaign 2 route	1	1	yes	144	1	144	0	2
2.2	with catalytic Pd	1	1	yes	139	3	139	0	3
2.3	with catalytic other metal	1	1	yes	124	9	124	0	5
2.4	nonmetal route	1	1	yes	132	6	132	0	4
2.5	with oxidation first	0	1	no	134	4	58	-78	n/a
2.6	nitro thiol route	1	1	yes	93	35	93	0	n/r
Alternative Leaving Group Variations									
3.1	diazonium LG	1	1	yes	102	25	106	+4	n/r
3.2	tosylate LG	1	1	yes	119	12	119	0	n/r
3.3	nitro LG	0	1	no	143	2	114	-29	n/a
3.4	nitrile route	1	1	yes	130	7	146	+16	1
Selected Other Route Options									
4.1	quetiapine-type best route	1	0	no	134	4	124	-10	n/a
5.4	tosyl chloride and aniline	0	1	no	111	16	100	-11	n/a
5.6	typical tosyl chloride based route	1	1	yes	104	23	111	+7	n/r
5.9	sulfobenzoic acid and aniline	0	1	no	129	8	112	-17	n/a
6.2	tolyl grignard with Protected Amide	1	1	yes	103	24	105	+2	n/r
7.2	protected aniline Grignard	1	1	yes	107	19	108	+1	n/r
8.3	sulfinic acid (lowest scored route)	??	1	??	70	38	71	+1	n/a
9.1	best bis-aryl sulfone route	1	1	yes	118	13	114	-4	n/r

^a n/a = not applicable; n/r = not ranked.

Obviously, using a subjective MCDM will have significant limitations. Agreement cannot always be reached on how to score particular issues and not all perspectives are equally correct. In an effort to make the MCDM process less subjective, Wu *et al.* (2009) proposed an alternative representation in form of a process flow diagram (Fig. A1.1). Although this approach helps to define how the information flows, it does not improve how the decisions are made, which will ultimately limit the range of applicability of the MCDM methodology.

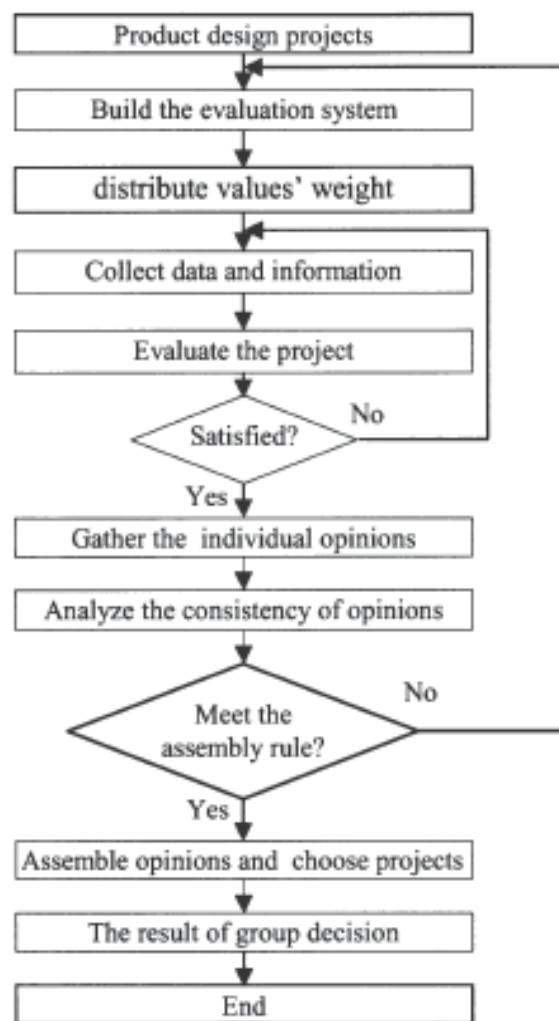


Figure A1.1: A Decision Analysis Flow Diagram for an MCDM (Wu *et al.* 2009)

A1.2. Ontology

Venkatasubramanian *et al.* (2006) proposes using ontology to design chemical products. Ontology is one of several tools developed over the past 20 years to support chemical product and process development and, in particular, it is used to explicitly describe information in terms of conceptual domains and their inter-relationships (Gruber, 1993). As a specific form of informatics, ontology describes classes and attributes on a case by case basis and can be used to combine information sources from a variety of databases in multiple languages. The process involves the development of both a syntax (i.e., structure) and the semantics (i.e., meaning) of the system. The end result should be able to automate tasks such as acquisition, representation, storage, manipulation, modification of, and reasoning with, data, information, and knowledge (Venkatasubramanian 2009). Although this is a long term view of the roles of ontology, the immediate benefit comes from the development of framework which enables optimization and design to proceed.

It has already been shown how decision analysis (e.g. MCDM) can be used to create and solve product design frameworks. Ontology represents a more rigorous mathematical approach to creating and solving chemical product design problems, including the identification of the various categories of information, followed by the construction of the information framework using an evolutionary design process consisting of proposing, implementing, and refining product classes, and then solving the design using automated decision analysis (Noy and McGuinness 2001). For example, the pharmaceutical industry utilizes information from molecular structure, biological interactions, solid dosages, batch manufacturing processes, Federal Drug Administration (FDA) controls, among many others. Terabytes of information are gathered and combined from these sources of information and combined into a large chemical

database resulting in tens of thousands of pages of reports (Venkatasubramanian 2009). To manage this onslaught of data Venkatasubramanian developed the Purdue Ontology for Pharmaceutical Engineering (POPE) to facilitate the screening of information, improve the decision-making, and create data linkages using intelligent systems engineering (Venkatasubramanian *et al.* 2006). In Fig. A1.2, an example of a ontological framework is presented using POPE. Knowledge is classified in three important categories: process knowledge, mathematical knowledge, and process control knowledge.

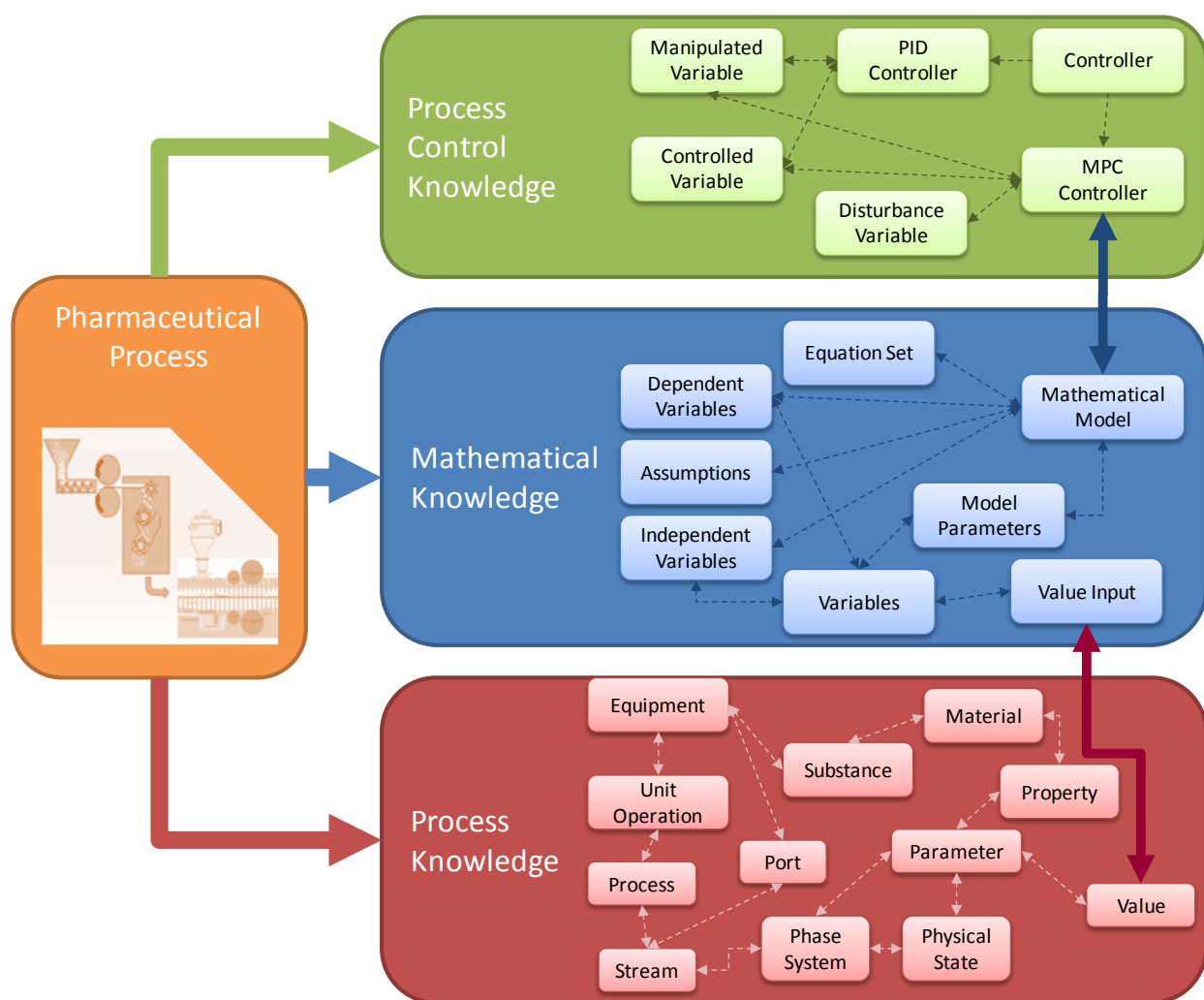


Figure A1.2: Ontological Relationships for a Pharmaceutical Process (Venkatasubramanian 2009).

To perform a chemical product design, the relationship between mathematical knowledge and process knowledge, specifically the components related to unit operations, material properties, and physical constructs are mapped into an evolutionary decision framework. Candidate solutions are then found by setting the desired values or important characteristics and executing the automated decision analysis to find the chemical products and processes that meet those conditions. Like all data driven models, the design is only as good as good as the database used to develop it. However the documentation requirements for heavily regulated industries like the pharmaceutical industry ensure enough information is captured to build a robust ontology decision network for chemical product design.

A1.3. Mathematical Optimization

As discussed in Appendix A1.1, the simplest way to choose the final product is to select one from a pre-screened set of options using MCDM. The MCDM method incorporates multiple important criteria (e.g. attributes and physico-chemical properties) into the decision-making process while providing a structured framework for the reduction of difficulty in complex decision-making (Wang *et al.* 1997). In the first step, candidate molecules are generated using combinatorial techniques and constrained using a set structure targets. Next, the potential candidates are then uploaded into the rows or columns of the MCDM matrix. Pertinent physico-chemical properties and/or attributes are loaded vice versa into the columns or rows of the MCDM matrix along with any additional criteria of importance, such as technical, economic, or environmental measures chosen using principles based on systematic analysis, consistency, independency, measurability, and comparability (Kepner and Tregoe 1997). In some cases, statistical tools such as predicted variance, min-max deviation, and correlation coefficients are used to select the best criteria to use in the MCDM. After the product criteria and potential

candidate solutions have been uploaded and normalized, a score is calculated for each candidate by multiplying the criteria's weighted importance by its normalized criteria value for each candidate and then summing across all criteria. The candidate with the highest score is considered "optimum." More details concerning this method are available in Appendix A1.1.

Alternatively, mathematical optimization is a completely automated method using only the constraints provided in the formulation with no subjective weighting of solutions. In comparison to MCDM that finds the best from a set of potential candidates, mathematical optimization finds the best from *all* feasible candidates. Thus the mathematical optimization is a more rigorous solution strategy, but comes at the cost of computational efficiency.

In most cases, the mathematical optimization involves finding the best molecular structure, from atomic backbone to long-range order morphology, given chemical structural constraints and objectives. The computational efficiency of a program is greatly enhanced when all of those constraints and objectives in the formulation are linear and continuous; a condition that is commonly referred to as Linear Programming (LP). Linear programs are solved using the simplex algorithm which finds the global optimum by searching the vertices of a solution space that is defined by linear target constraints (Biegler and Grossmann 1997). When binary variables are introduced, such as in Eqs. 4.2, 4.6, and 4.7, the problem becomes a Mixed Integer Linear Programming (MILP) problem. The simplest way to solve these problems is to use a direct brute-force method that enumerates each potential solution of binary variables, selecting the best as the global optimum. Unfortunately, this approach is computationally inefficient. A more elegant method is the branch-and-bound technique which works by relaxing and reformulating the problem so that no binary variables exist, as shown in Fig. A1.3. In this example, an objective function ($F_{obj}=Z$) is maximized subject to a set of constraints. Under relaxation, the

example objective value is found to be 41.25. Next, any y variables (e.g. attributes or physico-chemical properties) that contain non-integer values are then set, one at a time, to integer values, creating a series of sub-problems known as nodes (Borchers and Mitchell 1994). Each sub-problem LP is solved and the nodes that contain a higher objective function value are retained while those containing the lower values are cut. The process continues until there is no further improvement in the objective function (Biegler and Grossmann 1997). The performance of the branch-and-bound algorithm depends heavily on the choice of the non-integer variable to be used for branching and the selection of the node to be branched. As a rule, the noninteger variable to be chosen should be the one with the largest fraction (Gurdal *et al.* 1999), noted as x_2 in the example reformulation. The determination of which node to branch should be the one with the smallest objective value, which should generate a feasible design with a tighter upper bound. Nodes with objective functions higher than the relaxed value (Node 4) are deemed infeasible (i.e. beyond the relaxed variables solution domain). A branch-and-bound algorithm workflow, shown in Fig. A1.4, was first proposed by (Land and Doig 1960) and later refined by Borchers (1992).

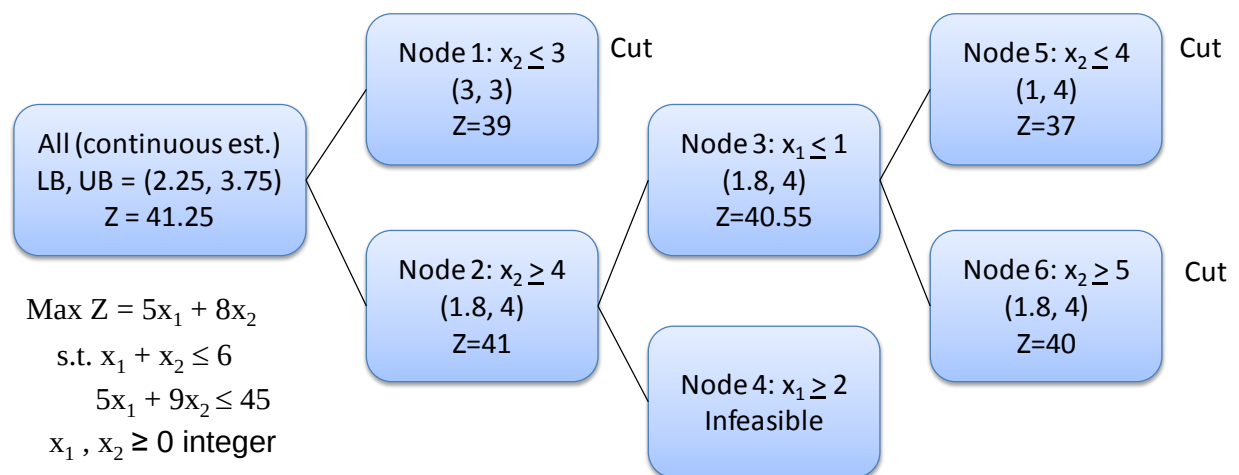


Figure A1.3: A Branch-and-Bound Example with an Optimum Solution Found Using Integer Programming within the Boundary Domain (Borchers and Mitchell 1994).

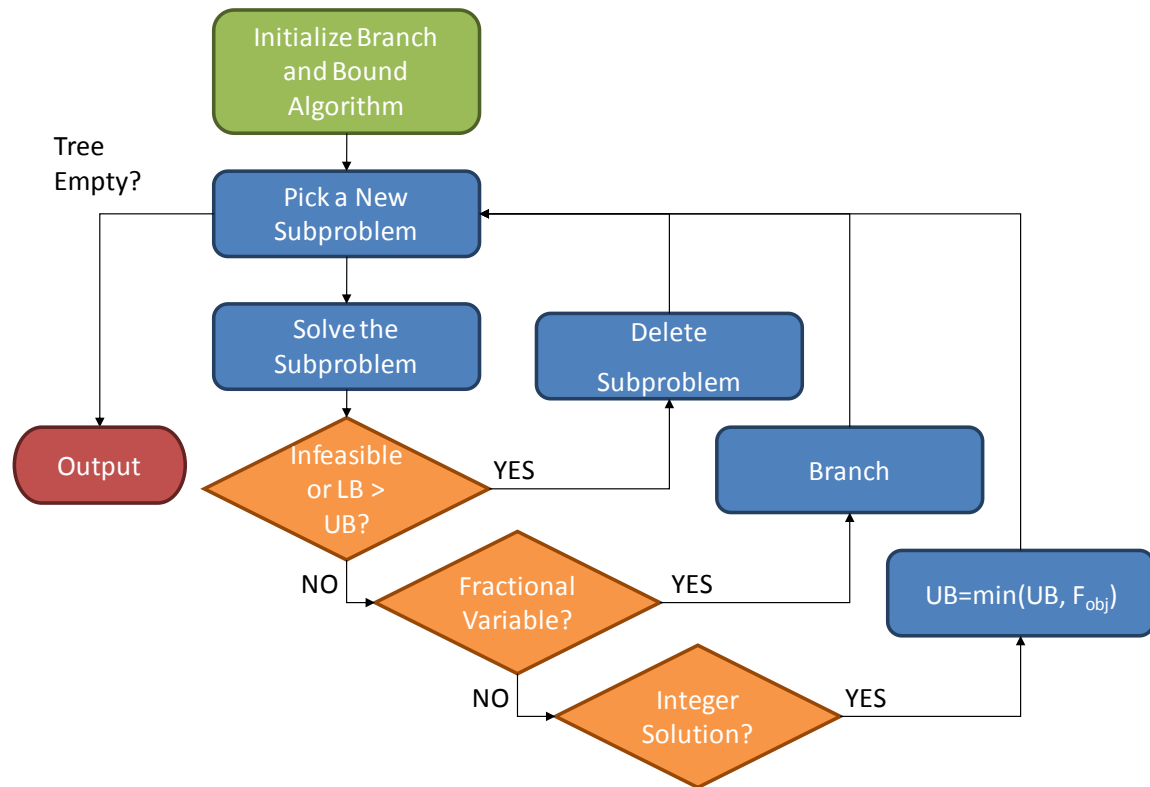


Figure A1.4: A Branch-and-Bound Algorithm Workflow (Borchers 1992).

The branch-and-bound algorithm has also been adapted for Non-Linear Programming problems (NLPs). In NLPs, the objective functions and constraints may be linear and/or nonlinear and such problems can be solved in a variety of ways. The easiest is to reparameterize the problem as a sequence of linear programs (Schmit and Farshi 1973), commonly referred to as Sequential Linear Programming (SLPs), or reduced gradient methods. Using a Taylor series expansion is the most common reformulation method in this construct, and it is solved by standard methods such as simplex optimization or interior point methods. It is most efficient for problems with a large number of linear constraints and well described model of nonlinear solution topography (e.g. a response surface with many valleys). However, in many cases, a Taylor series expansion does not match the curvature of the reformulation well enough and is, therefore, inappropriate to use (Gockenbach and Symes 2003). In this case sequential quadratic

programming (SQP) methods, which reformulate the problem into an alternative solution space using a Lagrangian of the objective function, are utilized. SQP is one of the most popular algorithms for nonlinear continuous optimization because of its robustness. Within the Lagrangian is the second partial derivative matrix, or Hessian function, which is used to develop a series of quadratic problems that result in fewer calculations and faster solution of the problem, especially in comparison with the reduced gradient method (Biegler and Grossman 1997, Sammons 2009). Although both of these tools are powerful methods for solving NLPs, neither one can guarantee that the solution they find is indeed the global solution unless the problem is completely convex and differentiable (Biegler and Grossmann 1997; Sammons 2009).

Finally, the most complicated optimization problem is the Mixed Integer Non-Linear Programming (MINLP) problem which utilizes both integer and non-linear programming techniques. This approach makes use of the branch-and-bound algorithm with special controls on the reparameterization of both the objective and constraints. The primary methods used to solve MINLPs are Outer-Approximation (OA) and Generalized Benders Decomposition (GBD) (Biegler and Grossman 1997, Sammons 2009). The solution strategy of both of these approaches is presented in Fig. A1.5.

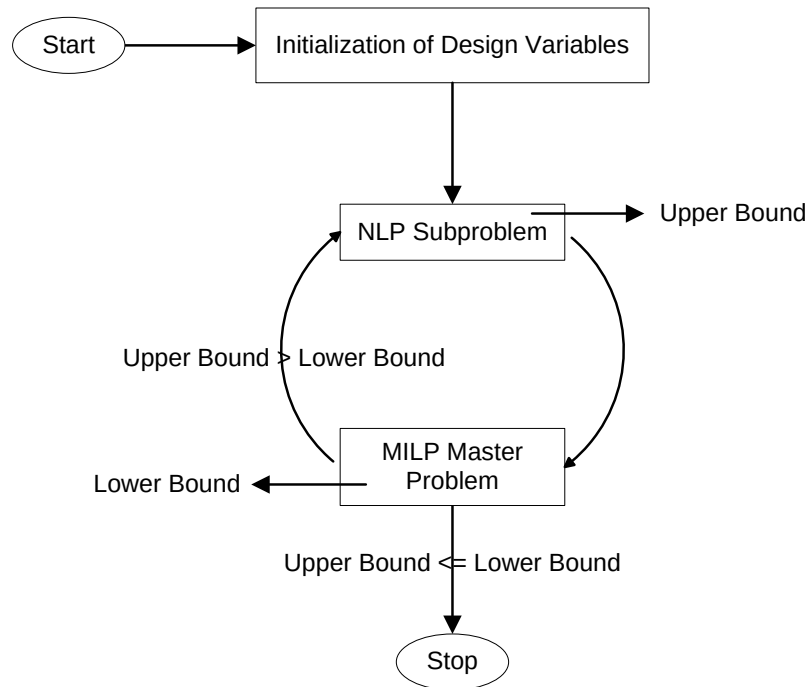


Figure A1.5: Flowsheet of MINLP Optimization (Diwekar 2003; Sammons 2009).

After initialization, the OA method uses an alternating sequence of NLP and MILP sub-problems with the MILP providing fixed binary variables to the NLP, which then determines new bounds for the MILP. The process continues until no lower bound is found below the current best upper bound. The GBD is very similar, only using the largest Lagrangian approximation obtained from the NLP sub-problem in the MILP sub-problem (Biegler and Grossmann 1997; Sammons 2009). However, since both methods use a NLP in the solution to the MINLP, a global optimum solution cannot always be guaranteed. Some improvement has been made using alpha branch and bound techniques (α -BB) (Sahinidis 2004), which heavily constrain the problems. Nevertheless, as these techniques are applied to increasingly larger problems that begin to mimic real world problems, the computational expense and ability of the methods come into question. In response, the trend has been to combine these deterministic methods with tools that serve to smooth the response surfaces of the solution domain in an effort to prevent the trapping of the

global optimization techniques in local minima solutions. These hybrid approaches most often make use of stochastic tools to loosely guide the more exact deterministic techniques into an area of solution.

Most hybrid methods incorporate one of two types of stochastic search techniques, Simulated Annealing (SA) and Genetic Programming (GP). These tools are especially adept at finding solutions where a global minimum is sought, like for programs involving nonlinear functions which may have multiple local minima. Both of these algorithms mimic naturally occurring phenomena and their implementation utilizes statistical tools associated with random selection and probabilistic decisions (Gurdal *et al.* 1999). Simulated annealing is a technique using that uses a probabilistic step function to perturb the solution in an effort to prevent the trapping of the algorithm in a local minimum. The perturbation is usually constrained such that the move is only adopted if the new solution has a better value for the objective function.

Genetic algorithms are techniques derived from biology that approximate the evolutionary principles of Darwin, mimicking the mechanism of natural selection by applying reproduction, crossover, mutation, and permutation to iterative solutions found in programming. These methods do not work in a sequential manner, but rather jump from point to point in the solution space, which keeps many solution points that may have the potential of being close to minima in the pool during the search process, rather than risk converging on a single point that may not be the global optimum (Gurdal *et al.* 1999). As shown in Fig. A1.6, in the first step of a genetic algorithm, an initial population is generated using a probabilistic function meant to offer good cross section of the design space to be investigated. The solutions in the initial population are also measured by their fitness, which is the value of the objective function in unconstrained problems. For constrained problems, such as in chemical product design, the fitness measure

must also take into account the proximity of the solution to the constraint margins. Based on the fitness measurement of the solutions in the population, the genetic algorithm can elect to apply a filter to make changes to the population for the next generation. For instance, if the reproduction filter is chosen, a probability function is applied where solutions with improved values of the objective are chosen more often than those whose objective functions are poor. A crossover filter is the swapping of solutions of two parent populations to create two child populations in the next generation and mutation involves randomly changing some of the solutions to be swapped. A permutation filter, also referred to as immigration, is the adding of new solutions into the population that have little similarity to the existing solutions in the population. Once these filters are applied, a new generation of solutions is measured by their population fitness, and the process is repeated. Termination of the genetic algorithm occurs when a set number of generations have been evaluated, or an approximate solution in which no noticeable improvement occurs in subsequent generations has been discovered (Diwekar 2003; Sammons 2009). Once these conditions are met, a constrained branch and bound deterministic optimization can be conducted.

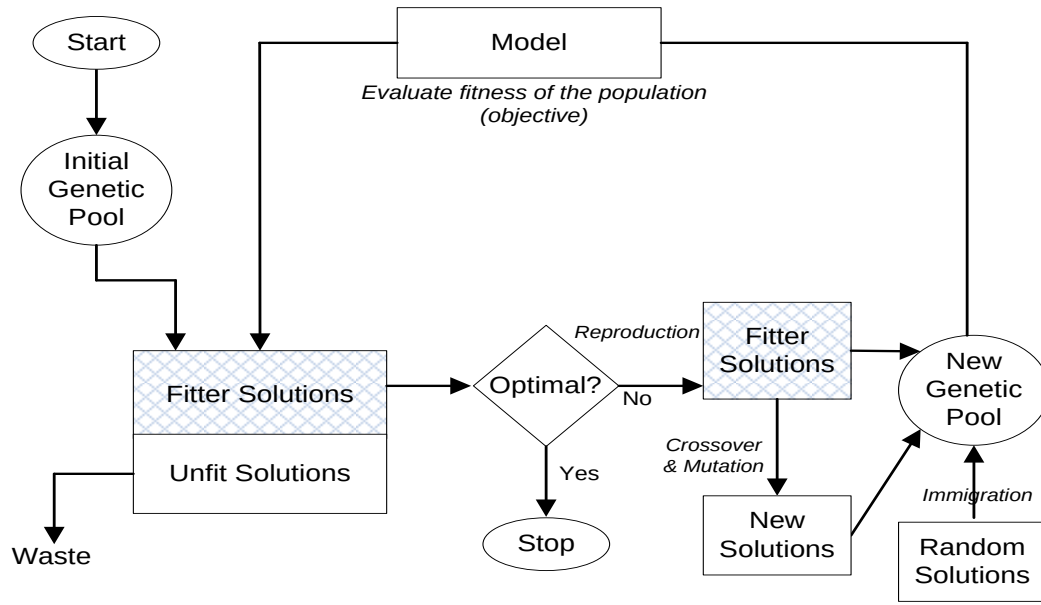


Figure A1.6: Genetic Algorithm Work Flow (Diwekar 2003; Sammons 2009).

Although stochastic search techniques can help center deterministic branch-and-bound optimization algorithms near the global optimum, they are computationally expensive, requiring thousands of analyses. The final computational cost is primarily the result of the size of the design and the way constraints hinder movement in the design space (Gurdal *et al.* 1999). They are also dependent on the requirements regarding reliability. It is common that each execution of the algorithm can result in a different solution. How close the solutions consistently get to the global optimum is the reliability measure. The goal is to have the stochastic search technique “practically equal” to the global optimum so that the optimum found by the deterministic branch-and-bound technique is the global optimum (Le Riche and Haftka 1995; Gurdal *et al.* 1999).

Appendix A2

Additional Characterization Techniques

This section contains summaries of other characterization techniques that may be used in chemical product design to generate the **P** property structure descriptor data. Many useful techniques exist for determining the molecular architecture of chemical product. Mass spectrometry provides information on the molecular formula and size of the chemicals in the product. Chromatography techniques provide more information on the polarity and size of the specific components of the product. Spectroscopy such as IR, UV-Vis, and NMR 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 product. Other characterization techniques provide information at larger length scales. Scattering and diffraction techniques provide information on particle size, morphology, and crystal packing. For polymeric structures, thermomechanical techniques like torsional braid analysis, dynamic mechanical analysis, thermogravimetric analysis, calorimetry, and differential thermal analysis are used to investigate phase transitions and reaction conditions in polymer chemistry. Viscoelastic properties are investigated using a variety of rheometric tools.

A2.1. Ultraviolet - Visible Spectroscopy

Ultraviolet-visible (UV-Vis) spectroscopy is only applicable to conjugated organic and metal ion systems and, hence, is less common than other spectroscopy methods (McMurry

1995). It measures the amount of energy needed to promote an electron from one orbital to another in a system of connected p-orbitals. The amount of energy needed can be directly related to the nature of the conjugated system. Similar to IR spectroscopic systems, it uses the Beer-Lambert law to relate the absorbance of light to the presence of specific conjugated systems within the structural groups of the chemical product. Although it is useful in distinguishing between enantiomers, its application is beyond the scope of this dissertation.

A2.2. Nuclear Magnetic Resonance Spectroscopy

Both vibrational and UV-vis spectroscopy provide information about functional group structure of the molecular architecture of chemical products. An alternative spectroscopic technique called Nuclear Magnetic Resonance (NMR) can also provide this information but is also capable of characterizing materials in dynamical states and slightly larger length scales. This property makes NMR extremely useful for studying structure-property relationships of biological systems. In general, each NMR signature depends on the magnetic environment of the NMR active nuclei, providing unique information about polymers on an atomic-length scale (Cheng and English 2003). The signatures of the NMR spectra can be assigned to specific atoms along the polymer backbone and side chains (Bovey 1982; Bovey 1988). The technique can be used to monitor the extent of reaction, check the purity of polymers, identify unknown materials, and to study polymer microstructure, dynamics, and interactions (Cheng and English 2003). The spectra may also contain information about the polymerization mechanism, side reactions, compositional heterogeneity, and morphology.

NMR is an absorption spectroscopy involving the absorption of radio frequency electromagnetic waves (Atkins 1998). Essentially, an atom is placed in an electromagnetic (EM) field and the change in the spin state of the nucleus of the atom is measured. The nuclei must have an

odd number of protons and neutrons to exhibit spin. ^1H , ^{13}C , ^{19}F , ^{15}N , and ^{29}Si are of the most interest. The signal strength of ^1H (or proton NMR) is approximately 50 times larger than ^{13}C due to the relative abundance of ^1H and is therefore the preferred absorption spectra. From quantum mechanics, when the nuclei are placed in an EM field, they adopt specific orientations either “parallel to” or “antiparallel to” (Mirau 1993). The two orientations do not have the same energy, with the parallel orientation typically having a slightly lower energy making it a slightly favored orientation (McMurry 1995). When the nuclei are subjected to EM radiation at the right frequency or resonance, the lower-energy state “spin-flips” to the higher energy state. The frequency necessary for resonance depends on both the strength of the magnetic field and the identity of the nuclei. Resonance frequencies in terms of empirical quantities are called chemical shifts, which are related to the difference between the resonance frequency of the nucleus and that of a standard reference, thereby ensuring that the chemical shift is independent of the frequency of the spectrometer (McMurry 1995). The standard typically used is tetramethylsilane (TMS). Likewise, the NMR absorption peak for a given nucleus is also directly proportional to the number of these atoms in the sample. Much like vibrational spectroscopy, the chemical environments in which these nuclei exist are often the same, leading to a classification scheme based on group theory. Developing the ‘NMR groups’ would be a worthwhile task, although it is beyond the scope of this dissertation.

NMR can also be extended beyond group theory to include polymeric information. Two major types of polymer NMR exist: solution and solid. Solution NMR evaluates the primary chemical structure of the macromolecular chain. Solid NMR, though more difficult to achieve a signal response, provides information on the primary structure, secondary structure (e.g. 2D ordering such as hydrogen bonding between chains), and the tertiary structure such as globular

folding in proteins or the three dimensional crystal structure. Solution NMR is the preferred method for naturally occurring polymers and most thermoplastic synthetic polymers, but is limited to solutions, gels, dispersions, melts, etc. Solid NMR is required for most thermoset synthetic polymers, even for primary structure information. Other modifications to the NMR experiment include frequency domain (HR-NMR), time domain (LR-NMR), liquid chromatography (HPLC-NMR & LC-NMR), size exclusion (SEC-NMR), supercritical (SFC-NMR), and electrophoresis NMR (Cheng and English 2003).

A2.3. Mass Spectrometry

Mass spectrometry is an analytical technique that elucidates the molecular structure of a molecule by measuring the mass-to-charge ratio for an individual molecule. Samples are first ionized and transported by magnetic or electric fields to an analyzer that sorts the ions according to their ratio.

$$\left(\frac{m}{Q}\right)a = E + v \times B \quad (\text{A2.1})$$

Where m is the mass of the ion, Q is the ion charge, a is the acceleration, E is the electric field, $v \times B$ is the vector cross product of the ion velocity and the magnetic field. The individual ratios of the atomic and molecular ions in the sample can provide an excellent estimation of molecular formula. To obtain more specific information of the molecular architecture, mass spectrometry is most often used in conjunction with gas, liquid, or gel permeation chromatography.

A2.4. Chromatography (Gas, Liquid, or Gel Permeation)

Chromatography is a characterization technique that relies on the differential affinity of a substance for a specific medium in order to indentify and quantify systems with complex

molecular architecture. The technique works by first introducing samples into a solvent, which carries them through column packed with materials specific to the type of separation desired. Separation types include observing the absorption affinities based on polarity (normal and reversed phase), size (gel permeation), charged ion site binding (ion-exchange), and molecular forces such as van der Waals, etc. (bioaffinity). The species are then quantified using a variety of detection mechanisms, one of which is mass spectrometry. Comparing the resulting chromatograph from a mass spectrometer with known samples in the same material column can elucidate differences in the molecular architecture. Other techniques, including light scattering could also be used.

A2.5. Scattering and Diffraction Techniques (Light, Neutron, X-ray)

Scattering is the process where some forms of radiated particles or waves are forced to deviate from a straight trajectory by one or more localized non-uniformities. The amount of scattering is proportional to the particle size or packing arrangement of the chemical product. In general, scattering implies interaction of waves with spatially uncoordinated (unordered) atoms. For small particles with wavelengths much smaller than the incident radiation, elastic light-scattering techniques like Rayleigh scattering can be used to measure both molecular weight and particle size. In this technique, the incident light intensity can be parameterized as a function of the scattering angle of the reflected light. The scattering angle θ is a measure of the bending of the radiation wavelength around a particle, which can be very useful for size characterization.

$$I = I_o \frac{1 + \cos^2 \theta}{2R^2} \left(\frac{2\pi}{\lambda} \right)^4 \left(\frac{n^2 - 1}{n^2 + 2} \right)^2 \left(\frac{d}{2} \right)^6 \quad (\text{A2.2})$$

Where R is the distance to the particle, θ is the scattering angle, n is the refractive index of the particle, and d is the diameter of the particle. Larger particles with wavelengths on the same order as the incident radiation can be described by Mie scattering. Particles with size much larger than the incident radiation wavelength are described by Geometric scattering.

Some of the most powerful scattering techniques use x-rays as the incident radiation. X-ray scattering analyzes the oscillations of electrons and nuclei by measuring the elastic scattering angles of incident X-rays. It is useful in obtaining information about the size, shape, and surface-to-volume ratio of macromolecules and lamellae in the nanometer to micrometer range. Types of x-ray scattering include small angle x-ray scattering (SAXS), wide-angle x-ray scattering (WAXS), and x-ray reflectivity. As the names of the techniques suggests, SAXS determine molecular architecture on the nano to micro length scales, which results in a scattering angle near 0 degrees. WAXS evaluates angles larger than 5 degrees, determining the molecular architecture at the molecular and nano-scales. Unlike many other characterization tools, scattering techniques can be used to study structural features present in materials at length scales that vary from angstroms (10^{-10} m) to microns (10^{-6} m) (Tolle *et al.* 2009).

Similar to scattering techniques, diffraction can also be used to determine molecular architecture from incident x-ray bombardment. Diffraction is in fact a combination of scattering and interference. The interference results from systems with ordered or packed periodic molecular architecture. Hence, diffraction is almost universally used to study crystalline solids whereas scattering can be used on liquids and amorphous solids. Diffraction characterization measurement techniques measures both the part of the wave that strikes an object and how the wave that passes through the object begins to occupy the space vacated by the blocked wave. Types of x-ray diffraction techniques includes single-crystal x-ray diffraction, which determines

the complete structure of large crystals, powder diffraction (XRD) which determines the grain size and preferred orientation, and thin film diffraction, used to characterize the crystallographic structure and preferred orientation of substrate anchored films. If the molecular architecture of the product is that of single crystals then single crystal x-ray diffraction is ideal for determining the crystal structure (Stephenson *et al.* 2001). More common is that the molecular architecture is in powder form, with many crystals, which calls for XRD methods to determine its structure. For XRD, the wavelengths of the diffracted beams λ are related to interplanar spacings d in the crystalline powder according to a mathematical relation called Bragg's Law (Connolly 2007):

$$n\lambda = 2d \sin \theta \quad (\text{A2.3})$$

Where n is an integer and θ is the diffraction angle. The diffraction angle, usually reported in 2θ form, explains many elements of crystalline structure. X-rays can either scatter uniformly from the atomic centers, which indicates atomic structure, or non-uniformly from the packed arrangement of atoms, indicating a specific type of crystalline structure (Connolly 2009). Since a large fraction of materials scatter isotropically (Bragg Diffraction), the scattering intensity can be reduced to a single one-dimensional spectra q as a function of the scattering vector (Tolle *et al.* 2009),

$$q = \frac{4\pi \sin \theta}{\lambda} \quad (\text{A2.4})$$

Areas of interaction lead to varying intensity responses that result in diffraction patterns, which are then interpreted in terms of molecular architecture (Atkins 1998). XRD probes the orderly arrangement of molecules in a the crystal lattice of solids and is more dependent on long range order effects and to lesser degree, the determination of the degree of crystallinity (Stephenson *et*

al. 2001). Two primary techniques exist, one using the individual peaks, the other using the whole pattern intensity.

In some cases, inelastic scattering occurs. Forms of inelastic scattering include Raman scattering and, in some cases, some types of x-ray scattering and diffraction. Inelastic scattering means that the kinetic energy of the incident radiation is not conserved, and changes in this energy are dependent on the molecular architecture. Raman scattering measures this as a shift in the wavelength frequency of the photon and is discussed in more detail in a previous section. When x-rays are scattered inelastically and lose energy, they are exhibiting Compton scattering. More information on this technique can be found in Klein (2002) and Atkins (1998).

A2.6. Other Characterization Techniques

Beyond spectroscopy related characterization are the thermo-physical tools for characterization. These are loosely based on the concept of calorimetry which measures the temperature change of a sample against a control under induced heat transfer, chemical reaction, or applied physical stress. A standard solution calorimeter describes heats of reaction, mixing, solution, and dilution. When combined with a differential scanner that increases the temperature at a set rate, the calorimeter can provide additional information on the phase transitions of polymers, including the glass transition temperature T_g , crystallization temperature T_c , and the melting point temperature T_m , by measuring the changes in heat flow into the system. The glass transition temperature is the point at which free rotation of bonds ceases and the polymer becomes frozen in an entangled, disordered state (Rodriguez *et al.* 2003). Above this temperature segmental motion begins to occur. The crystallization temperature is the point at which the polymer frees itself enough to begin realigning itself in a lower energy crystalline state. The melting temperature is the highest point in which a crystal lattice is stable.

Temperatures above T_m result in endothermic melting. Similar to the differential scanning calorimeter (DSC) approach, differential thermal analysis (DTA) also provides information on the long-range order of thermo-physical properties of chemical products. As opposed to DSC, it works by administering a steady heat transfer rate and measuring a temperature difference. Both of these techniques are useful for long range order structural differences when the structural similarity renders diffraction techniques difficult to interpret (Atkins 1998; Rodriguez *et al.* 2003).

Another useful thermo-physical technique is dynamic mechanical analysis (DMA). This technique is used to measure the mechanical properties of a wide range of materials, specifically the viscoelastic properties and loss moduli of polymers as a function of temperature and frequency (time). The technique works by applying sinusoidal stress (or strain) and measuring the reciprocal strain (or stress). Likewise, the temperature of the sample is often varied, which allows for the investigation of the phase transition temperatures. Types of viscoelastic properties that can be described are the peak tensile stress σ , dynamic (Young's) modulus E , often reported as a combination of elastic storage modulus E' and the dissipated loss modulus E'' , and dynamic shear modulus G , reported as storage modulus G' and loss modulus G'' . These mechanical properties are extremely important in the physical handling of the chemical product during manufacturing, as well as when the chemical product is used in a structural configuration.

Torsional braid analysis is another method of evaluating the thermomechanical behavior of the chemical products that cannot support their own weight. In this technique, the chemical product is placed into a glass braid, which is then oscillated under increasing temperatures. Changes in the oscillation behavior can be directly related to the viscoelastic properties of the sample. In a similar technique, Thermogravimetric analysis (TGA) measures changes in

molecular weight of chemical products as a function of measured temperature changes, looking specifically for the presence of adsorbed moisture, solvents, or other entrapped contaminants.

Thus far the majority of these viscoelastic techniques discussed are looking for long-range order structural changes of chemical products based on temperature changes. In contrast, rheometry uses rheological properties like viscosity to estimate thermomechanical behavior such as yield stress, modulus, creep, and recovery. Types of rheometric tools include capillary and rotational geometries.

Finally, microscopy is one of the most common and simple techniques used in characterization. Particularly useful are optical microscopy, scanning electron microscopy, and transmission electron microscopy (TEM). Optical spectroscopy uses light to illuminate structures that cannot be seen with the naked eye. It is particularly adept at determining the phase of single crystalline structures. SEM works by scanning a fine beam of high energy electrons over the surface of a sample which interacts to produce signal images of the surface's topography and composition. It typically has micron scale accuracy, making it ideal for determining polymer morphology and particle size. TEM is similar to SEM, but works on the smaller nanometer scale structures.

Appendix A3

Additional Decomposition Tools

This section contains additional decomposition tools that can be used to develop the latent variable structure of the chemical product.

A3.1. Network Component Analysis

Conventional approaches, such as PCA, typically seek a loadings matrix $\mathbf{L}$ such that the resulting scores matrix $\mathbf{T}$ satisfies orthogonality or independence criteria (Liao *et al.* 2003). When dealing with data generated from structured networks where some information pathways are known, and thus constrained, these decomposition techniques present two drawbacks (Liao *et al.* 2003; Tolle *et al.* 2009). First, the implicit statistical assumptions on the data structure, i.e. least squares regression, preclude the inclusion of physical and chemical systems understanding. Second, the reconstructed connectivity structure is unlikely to be consistent with the known underlying phenomenological structure. Therefore, a decomposition method is sought that makes no assumptions on the statistical properties of the system, but rather allows proper handling of prior knowledge on the structure characterizing a given system (Liao *et al.* 2003).

Tolle *et al.* (2009) proposes that Network Component Analysis (NCA) provides the best opportunity for integrating this *a priori* knowledge. For PCA, the solution is inherently nonunique, meaning that an infinite number of scores and loadings structures may be derived and the best chosen through least squares analysis. NCA constrains the loading structure to known

phenomenological bipartite topology, to which convex optimization can be applied (Tolle *et al.* 2009). This effectively places zeros in the loadings matrix where no established relationships exist, leaving non-zeros where they do. The problem can then be reformulated as an MINLP where the NLP is solved using the IPOPT algorithm for each combination of integers (Wachter and Biegler 2006; Tolle *et al.* 2009). When applied to process-structure-property relationships found in chemical products, NCA gives the data structure shown in Fig. A3.1. For example, Tolle *et al.* (2009) used network component analysis (NCA) to decompose wide angle x-ray spectroscopy (WAXS) data into a series of network connections of material structural units that could quantitatively describe the complex interactions occurring at this scale. The key feature of this approach is that unique, physically meaningful solutions can be identified with little or no *a priori* information regarding the system's characteristics as long as accurate and reliable characterization data is available (Tolle *et al.* 2009).

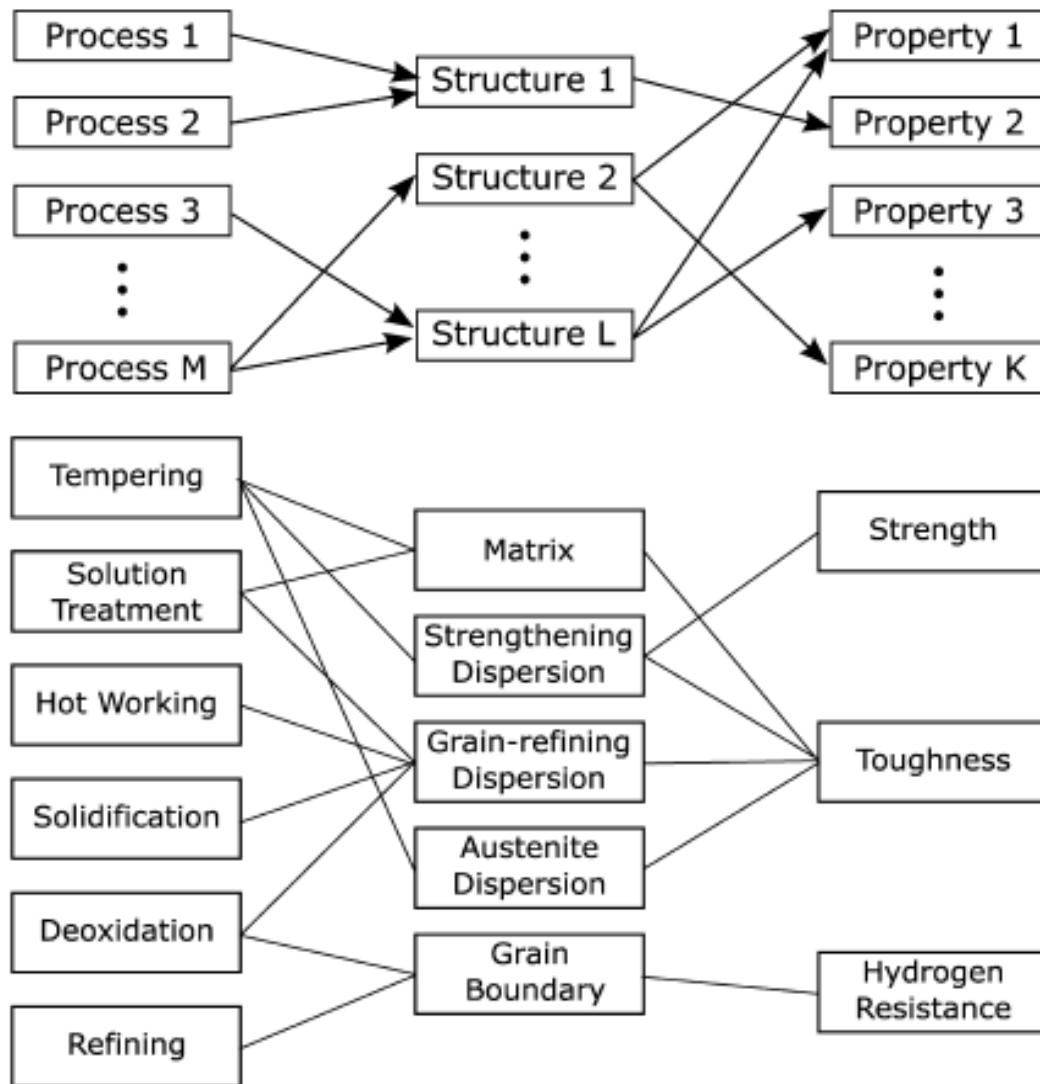


Figure A3.1: A NCA illustration of process-structure-property relationships for high performance alloy steel (Olson 1997; Tolle *et al.* 2009).

Appendix A4

Additional Physical-Chemical Property Models

This section contains alternative physico-chemical property models using either *ab initio* simulation techniques (e.g. quantum mechanics (QM), molecular mechanics (MM), molecular dynamics (MD), Brownian dynamics (BD), -dissipative particle dynamics (DPD), monte carlo (MC) stochastic techniques, and topological indices, etc.) and/or less accurate but more efficient topological indices (T-I-) combined with quantitative structure property (QSPR) models.

A4.1. Simulation Techniques

Recent years have witnessed an increase in the number of people using computational chemistry to estimate chemical and physical properties of systems (Young 2001). The advent of computational chemistry software has brought these powerful techniques to the forefront of chemistry and chemical engineering. Venkatasubramanian (2009) notes that the use of these techniques represents the coming of a “data deluge” in the chemical engineering field. The data is generated by either tracking the electrons (e.g. QM), molecules (e.g. MM, MD), or molecular clusters (e.g. BD, DPD, etc.) of the system. The development of chemical architectures at different scales will require a fundamental understanding of the hierarchical structure and behavior of these systems and will result in multiscale modeling and simulation strategies that provide seamless nesting and coupling (Zeng *et al.* 2008).

In most cases, traditional multi-scale design problems begin with *ab initio*, or “from the beginning” calculations. *Ab initio* calculations are computational chemistry tools that are derived directly from theoretical principles and work by tracking the location probabilities of the electrons in a molecular structure. *Ab initio* methods are based on a quantum mechanical (QM) description of electrons in an eigen equation, first proposed by Schrodinger and Heisenberg (Atkins 1998):

$$\hat{H}\Psi = \hat{E}\Psi \quad (\text{A4.1})$$

where $\hat{H}$ is the Hamiltonian operator, Ψ is a Schrodinger wave function of the electron and nuclear positions, and $\hat{E}$ is the energy. The Born interpretation states that the square of the wave function corresponds to a probability function describing its location (Atkins 1998). Dynamical information is provided by the Hamiltonian operator, which describes the electronic motion of the particles, both nuclei and electrons, in the system (Young 2001).

$$\hat{H} = - \sum_i^{\text{particles}} \frac{\nabla_i^2}{2m_i} + \sum_{i < j}^{\text{particles}} \sum \frac{q_i q_j}{r_{ij}} \quad (\text{A4.2})$$

where ∇_i^2 is the Laplacian operator on particle i , m_i is the mass of particle i , and q_i is the charge of particle i , and r_{ij} is the distance between particles. The first term describes the kinetic energy of the particle and the second term describes the Coulombic attraction and repulsion of particles (Young 2001). The calculation of chemical and physical properties from this derivation is achieved by first taking the expectation of the Hamiltonian operator and solving for the potential energy of the system (Iachello and Levine 1995; Young 2001).

$$\langle E \rangle = \int \Psi^* \hat{H} \Psi \quad (\text{A4.3})$$

Other observable, chemical and physical properties can either be substituted, or calculated from the potential energy surface using the Hellmann-Feynman theorem (Iachello and Levine 1995) which relates the derivative of property specific energy to the specific property Hamiltonian (Young 2001).

$$\frac{dE}{dP} = \left\langle \frac{\partial \hat{H}}{\partial P} \right\rangle \quad (\text{A4.4})$$

It should be noted that Eq. A4.4 is the time-independent, non-relativistic version of the Hamiltonian for the Schrodinger equation. Additional terms appear in the Hamiltonian when either time and position relativity or electromagnetic interactions occur (Young 2001). For instance, a more computationally friendly form of the Hamiltonian can be found using the Born-Oppenheimer approximation. This approximation adjusts the Hamiltonian to treat the heavier, slower moving nuclei as stationary in reference to the electrons moving around it (Atkins 1998):

$$\hat{H} = - \sum_i^{\text{electrons}} \frac{\nabla_i^2}{2} + \sum_i^{\text{nuclei}} \sum_j^{\text{electrons}} \frac{Z_i}{r_{ij}} + \sum_{i < j}^{\text{electrons}} \sum \frac{1}{r_{ij}} \quad (\text{A4.5})$$

where the first term is the kinetic energy of the electrons only, the second term is the attraction of the electrons to the nuclei, and the third term is the repulsion between electrons. Any motion of nuclei can be described by considering this entire formulation as a potential energy surface on which the nuclei can move (Young 2001).

Another common approximation is the Hartree-Fock (HF) which is the central approximation behind self-consistent central field theory (Atkins 1998). It addresses the difficult computation of electron-electron interactions by integrating the repulsion term to an average effect. The overwhelming benefit of this method is that it breaks the many-electron Schrodinger equation into many, simpler one-electron equations, which are then solved to yield individual

wave functions called orbitals (Young 2001). These individual orbitals represent the well-known electronic shell structure in chemical periodicity (Atkins 1998). The mathematical functions used to approximate the electronic wave function are called basis sets. The most common use a linear combination of Gaussian-Type Orbitals (GTO), represented as $\exp(-ar^2)$ (Young 2001).

One of the limitations of HF calculations is that they do not include electron correlation, which means this approximation tends to overestimate surface energy. Since this estimate is always equal to or greater than the exact energy, the HF can serve as a basis on which improvements in the mathematical functions of the approximations can be measured. A detailed account of other approximations can be found in various computational chemistry texts. In general, the relative accuracy is as follows (Young 2001):

$$\text{HF} \ll \text{MP2} < \text{CISD} \sim \text{MP4} \sim \text{CCSD} < \text{CCSD(T)} < \text{CCSDT} < \text{Full CI}$$

where MP2 and MP4 are the HF corrected using second and fourth order Moller-Plesset perturbation theory; CI and CISD are the corrections using the complete configuration interaction and the configuration interaction with single and double-excitation only; and CCSD, CCSD(T), and CCSD are the corrections using the coupled clusters at various orders of expansion. It should be noted that although these and other mathematical approximations can be used to simplify the computation, no experimental data or empirical models are used. Approximations are more prevalent in larger, more complex molecules, but are necessary because even the simple Hartree-Fock (HF) approximation scales to the N^4 basis function sets (Young 2001). Other approximations scale even worse, which limits the capability of this method to small to medium organic molecules.

A slight improvement in the molecular size modeling capabilities can be achieved using semi-empirical methods. Semi-empirical calculations utilize the structure of the HF

approximation with a Hamiltonian and a wave function, but replace some electron specific estimations with experimentally derived parameters, or decoupled approximations. A list of semi-empirical methods can be found in Young (2001). The advantage of these methods is that they are much faster, which allows their application range to include finding the geometry and dynamic energies of moderate sized molecules. However, semi-empirical methods give poor results for van der Waals and dispersion intermolecular forces due to the lack of diffuse basis functions, both of which are vital interactions in the meso-scale (Leach 2001).

Alternatively, density functional theory (DFT) has recently been shown to achieve similar accuracy to *ab initio* calculations while being computationally less intensive by estimating the energy of a molecule from its electron density instead of a wave function. However, because it is a relatively new method, the approach requires extensive validation which can limit the scope and size of the molecules computed.

Overall, the most severe limitation of *ab initio* calculations is the limited size of the molecules that can be modeled. Molecular Mechanics (MM) is a method that limits this computational intensiveness by calculating the energy of a molecule from a set of force field constants which are constant across all molecular types and can be derived from *ab initio* calculations or experimental data. This is best accomplished by using an atom based approach that implicitly includes electron information. In this approach, interaction potentials and their corresponding parameters, commonly known as the force field, describe in detail how the particles in a system interact with each other and by extension how the potential energy of a system depends on the particle coordinates (Zeng *et al.* 2008). Such a force field may be obtained by the *ab initio* quantum methods, empirical methods such as the Leonard-Jones, Morse, Born-Mayer, or quantum-empirical methods that nest quantum mechanics within

molecular mechanics such as the embedded atom model (Zeng *et al.* 2008). The criteria for selecting a force field include the accuracy, transferability and computational speed. A typical interaction potential U may consist of a number of bonded and nonbonded interaction terms (Zeng *et al.* 2008):

$$\begin{aligned}
 U(\vec{r}_1, \vec{r}_1, \dots, \vec{r}_n) = & \sum_{i_{bond}}^{N_{bond}} U_{bond}(i_{bond}, \vec{r}_a, \vec{r}_b) + \sum_{i_{angle}}^{N_{angle}} U_{angle}(i_{angle}, \vec{r}_a, \vec{r}_b, \vec{r}_c) \\
 & + \sum_{i_{torsion}}^{N_{torsion}} U_{torsion}(i_{torsion}, \vec{r}_a, \vec{r}_b, \vec{r}_c, \vec{r}_d) + \sum_{i_{inversion}}^{N_{inversion}} U_{inversion}(i_{inversion}, \vec{r}_a, \vec{r}_b, \vec{r}_c, \vec{r}_d) \\
 & + \sum_{i=1}^{N-1} \sum_{j>1}^N U_{vdW}(i, j, \vec{r}_a, \vec{r}_b) + \sum_{i=1}^{N-1} \sum_{j>1}^N U_{electrostatic}(i, j, \vec{r}_a, \vec{r}_b)
 \end{aligned} \tag{A4.6}$$

The first four terms represent bonded interactions: bond stretching, bond bending, dihedral angle torsion, inversion interaction. The last two terms are non-bonded interactions, van der Waals energy and electrostatic energy. Since electrons are not explicitly included, electronic property based processes cannot be modeled. The reduction in complexity afforded by MM calculations means that both large molecules and molecules with intermolecular forces can be modeled. The cost of this complexity reduction is that the method is extremely sensitive to parameterization, meaning that it is important that the basis sets be based on experimental and computational studies similar to the system being studied. Some of the most common basis set functions are as follows: assistant model building with energy refinement (AMBER), consistent force field (CFF), Carbohydrate hydroxyls represented by external atoms (CHEAT), generalized organic force fields (MM1-MM4), and the universal force field (UFF). Other force field basis sets and recommendations on when to use them can be found in readily available computational chemistry texts. In some cases it is advantageous to suppress hydrogen, choosing to model CH₂

as a group rather than as an sp^3 carbon bonded to two hydrogen atoms. This is known as the united atom force field and is common way to reduce the complexity of modeling large macromolecules.

Statistical mechanics provides the means to convert the individual optimized molecular structure of a single molecule into physical properties that describes many molecules in many conformations and energy states. Molecular Dynamics (MD) and Monte Carlo (MC) simulations are the most common methods for generating these relationships. MD simulations model the time-dependent behavior of molecular systems, such as vibrational or Brownian motion. MD are the time dependent form of molecular mechanics, solved in discrete time steps, under different ensembles of constraints. Examples include grand canonical (μVT), microcanonical (NVE), canonical (NVT), and isothermal-isobaric (NPT) (Zeng *et al.* 2008). It is important to validate that the simulation is delivering chemical properties that exhibit normal Gaussian distributions and that they are matched with experimental data since this type of time step simulation is highly susceptible to systematic error propagation. MC simulations differ from MD simulations in that the location, orientation, and geometry of a set of molecules is randomly chosen according to a statistical distribution (Young 2001). As a result, MC simulations are less computationally intensive than MD simulations, but are limited since they cannot yield time dependent information. Both of these statistical mechanics simulation methods can become as computationally expensive as *ab initio* calculations due to the large number of molecules being calculated, especially for large molecular systems.

Bridging the gap between these molecular methods and true continuum behavior has been the focus of recent research. Whereas the molecular interactions of small molecules can be easily described by existing thermodynamic models, large molecules, particularly those with

nano-, micro-, and meso-scale architectures still rely on combinatorial models to describe their behaviors. Common techniques include Brownian Dynamics (BD), Dissipative Particle Dynamics (DPD), Lattice-Boltzmann (LB), and time-dependent Ginsburg-Landau (TDGL), and dynamic DFT (Zeng *et al.* 2008). Brownian dynamics replaces the explicit definitions of small molecule solutions with implicit continuum definitions which allows for a much larger time step to be used. Interaction forces on individual atoms are estimated by the Langevin equation, summing effects from conservative forces, momentums, and random Gaussian noise. It should be noted that this approximation means that energy and momentum of the solution is no longer conserved, meaning that the macroscopic behavior is not hydrodynamic and does not obey the Navier-Stokes equation. Like MD and BD, DPD is a particle based method that can simulate fluidic behavior. However, it uses a coarse-grained molecular assembly as a basic unit instead of single atoms (Zeng *et al.* 2008). In this manner it parameterizes the interaction forces into conservative, dissipative, and random forces. The nature of the forces is pairwise and steps can be taken to ensure momentum is conserved, meaning that this method obeys Navier-Stokes. The details of these and other techniques can be found in Zeng *et al.* (2008).

A4.2. Topological Indices

The representation of a molecule in the form of a graph is the first step in the development of a topological index (Chemmanurathapillai *et al.* 2010). Using the molecular graph, a set of values for topological indices (TI) can be estimated. Various Quantitative Structure Property Relationships (QSPR) or Quantitative Structure Activity Relationships (QSAR) are then used to relate the TI values to a set of physical and chemical properties. The important qualities that make a meaningful TI include direct structural interpretation, correlation with at least one property, linearly independence, non-triviality, and a strong structural basis,

among others (Randic and Basak 2001). Although hundreds of topological indices can be found in the literature, only a few of them have found wide applications. Some of the more commonly used topological indices are connectivity indices, edge adjacency indices, shape indices, Wiener index, and the Hosoya topological index (Chemmanagattualappil *et al.* 2010). Of these, the most widely used topological index is the connectivity index (Trinajstić 1992).

The success of connectivity indices (CI) is due to the fact that each of its fragments can be calculated directly from the valence (Lewis) bond diagrams. Beginning with the atomic level, the first atomic index is the simple connectivity index, δ , equal to the number of non-hydrogen atoms to which a given non-hydrogen atom is bonded (Bicerano 1996). The second atomic index is the valence connectivity index, δ^v , which incorporates the electronic structure information of each non-hydrogen atom and is sometimes referred to as the valence delta.

$$\delta^v = Z^v - h \quad (\text{A4.7})$$

Here, the Z^v of an atom is the count of all adjacent bonded atoms and all π and lone pair electrons while h is the number of hydrogen atoms bonded to that atom. When dealing with high atomic weight atoms, the effect of non-valence electrons on the atomic size and properties must be considered (Keir and Hall 1986). For that, the valence delta is redefined for an atom of atomic weight Z as follows (Bicerano 1996; Chemmanagattualappil *et al.* 2010):

$$\delta^v = \frac{(Z^v - h)}{(Z - Z^v - 1)} \quad (\text{A4.8})$$

If i and j are the atoms involved in the bond, then the bond indices, β and β^v , are defined through the pairing of atomic indices (Kier and Hall 1986; Gani *et al.* 2005):

$$\beta = \delta_i \delta_j \quad (\text{A4.9})$$

$$\beta^v = \delta_i^v \delta_j^v \quad (\text{A4.10})$$

An alternative method for classifying these elements is to use the adjacency matrix, **J**, which originates in graph theory, and indicates which atoms are bonded (Ugi *et al.* 1979). From the adjacency matrix, the connectivity matrix, **C**, can be calculated. The connectivity matrix not only indicates which atoms are connected through covalent bonds, but also the number of covalent bonds between each atom. Further extension of this approach results in the bond-electron, **BE**, matrix which represents both the covalent bond and the free valence electrons of each atomic system (Ugi *et al.* 1979). Examples of these matrices for hydrogen cyanide (HCN) are given below (Ugi *et al.* 1979):

$$J(\text{HCN}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad (\text{A4.11})$$

$$C(\text{HCN}) = \begin{bmatrix} H & 1 & 0 \\ 1 & C & 3 \\ 0 & 3 & N \end{bmatrix} \quad (\text{A4.12})$$

$$BE(\text{HCN}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 3 \\ 0 & 3 & 2 \end{bmatrix} \quad (\text{A4.13})$$

Using the information from the atom and bond indices shown above, the calculation of the zero'th order (atomic) connectivity indices for the entire molecule can be defined in terms of the summation of the vertices of the hydrogen-suppressed graphs (Bicerano 1996):

$${}^0\chi = \sum_{\text{vertices}} \left(\frac{1}{\sqrt{\delta}} \right) \quad (\text{A4.14})$$

$${}^0\chi^v = \sum_{\text{vertices}} \left(\frac{1}{\sqrt{\delta^v}} \right) \quad (\text{A4.15})$$

The first order (bond) indices, ${}^1\chi$ and ${}^1\chi^v$, for the entire molecule are defined as the summation over the edges of the hydrogen-suppressed graphs (Bicerano 1996; Chemmangattuvalappil *et al.* 2010) are as follows:

$${}^1\chi = \sum_{\text{edges}} \left(\frac{1}{\sqrt{\beta}} \right) \quad (\text{A4.16})$$

$${}^1\chi^v = \sum_{\text{edges}} \left(\frac{1}{\sqrt{\beta^v}} \right) \quad (\text{A4.17})$$

Once the indices are built, various linear and non-linear regression models can be used to relate them to a specific physical property in QSPR and QSAR relationships. The general form of the relationship is listed below:

$$\theta = f(TI) \quad (\text{A4.18})$$

where θ is a property function corresponding to a property P . The total number and types of connectivity indices required for the models are dependent upon the properties dependence on particular types of interactions. In most instances, the first and second order connectivity indices are sufficient to describe the molecular graph. A finite number of higher-order connectivity indices have also been created to describe molecular interactions occurring at longer lengths. Examples of higher order connectivity indices include structural and topological information on ring substitutions, branches, non-hydrogen clusters, among others (Bicerano 1996).

Like the van Krevelen group contribution method (van Krevelen 1990) for the design of long range ordered systems in polymers, Bicerano (1996) has also developed similar CI methods

for polymeric backbones. In this method, Bicerano (1996) avoids the truncation errors associated with variable polymer chain lengths and the resulting molecular weights by basing the CI on an alternative representation of the monomeric graphs. Essentially, the graphs are built by completely enveloping one bond of the monomeric unit, and ignoring the other. Strong correlations for a variety of volumetric, thermodynamic, chemical, mechanical, optical, and magnetic properties have been shown to be related to topological indices (Bicerano 1996).

It should be noted that CI are more difficult to use in the design of chemical structures than the property models in the group contribution methods (GCM) (Chemangattuvalappil *et al.* 2009). In particular, GCM can be directly used in reverse problem formulations since their property functions can be expressed as linear expressions, opening up additional computational reduction possibilities. However, the applicability of group contribution methods is limited to a small number of properties and molecular groups. One way to address this problem is to broaden the application range of group contribution methods (GCM) by incorporating CI in an GC+ formulation (Chemangattuvalappil *et al.* 2009). It has been proposed that the first order connectivity indices can be written for particular groups. However, if first order connectivity indices for the different groups are written separately, their sum will not give the CI value of the molecule. This is because, the contribution to first order CI due to the bonds between two separate groups are not fully represented. Therefore, an additional term for the bond between different groups has to be included in the expression (Gani *et al.* 2005; Chemangattuvalappil *et al.* 2010):

$$({}^v\chi^1)_{group} = \sum_1^k \left(\frac{1}{\sqrt{\beta_{internalbonds}^k}} \right) + \sum_m \left(\frac{0.5}{\sqrt{\beta_{bonds\ out\ of\ groups}^m}} \right) \quad (A4.19)$$

Here, k is the number of bonds inside the group for which the expression is written and m is the number of free bonds in the group. The pure component property model proposed based on the CI model is given by Gani *et al.* (2005). Here, Y is the sought property, A_i is the number of atom i , a_i is the estimated contribution of atom i , while b , c and d are adjustable parameters.

$$f(Y) = \sum_i (a_i A_i) + b \left({}^v\chi^0 \right) + 2c \left({}^v\chi^1 \right) + d \quad (\text{A4.20})$$

The CI method for estimating pure component properties cannot produce very accurate results because, a large number of compounds are represented with very few parameters (Gani *et al.* 2005; Chemmangattuvalappil *et al.* 2010). So, its applicability is limited to the prediction of properties of the groups and/or property contributions which are not available in established datasets (Constantinou and Gani 1994; Constantinou *et al.* 1995; Marrero and Gani 2001).

$$f(Y_m) = \sum_i (a_{m,i} A_{m,i}) + b \left({}^v\chi^0 \right)_m + 2c \left({}^v\chi^1 \right)_m \quad (\text{A4.21})$$

$$f(Y^*) = \left(\sum_m n_m f(Y_m) \right) + d \quad (\text{A4.22})$$

$$f(Y) = \left(\sum_i N_i C_i \right) + f(Y^*) + \left(\sum_s N_s C_s \right) + \left(\sum_t N_t C_t \right) \quad (\text{A4.23})$$

where m is the number of different missing groups and n_m is the number of times the missing group is present in the molecule. $f(Y^*)$ is the property function of the missing group, A_{mj} is the property contribution of atom j , while a, b and c are the adjustable parameters estimated by Gani *et al.* (2005).

An alternative representation of CI is the use of molecular signature. Molecular signature is a molecular descriptor introduced by Visco *et al.* (2002) and Faulon *et al.* (2003a) that represents atoms in a molecule using extended valencies of a pre-defined height. If G is a molecular graph and x is an atom of G , the atomic signature of height h of x is a canonical

representation of the sub-graph of G containing all atoms that are at a distance h from x (Chemmanattuvalappil *et al.* 2010). Topological indices like CI generally only look one or two bonds away. Conversely, molecular signature can look an unlimited distance away, meaning more structural information can be captured, leading to much improved accuracy in the design of molecular architectures. The systematic procedure for the construction of atomic signature developed by Visco *et al.* (2002) is explained in more detail by Chemmanattuvalappil *et al.* (2010).

Once the atomic signatures of all atoms are estimated in this way, the molecular signature can be represented as the linear combinations of all atomic signatures. If ${}^h\sigma_G({}^hX_i)$ is a base vector, ${}^h\alpha_i$ is the number of atoms having the signature of the base vector and ${}^hK_G$ is the number of base vectors, then the molecular signature ${}^h\sigma(G)$ is represented as:

$${}^h\sigma(G) = \sum_{x \in V_G} {}^h\sigma_G(x) = \sum_{i=1}^{{}^hK_G} {}^h\alpha_i {}^h\sigma_G({}^hX_i) \quad (\text{A4.24})$$

These signatures can produce meaningful quantitative relationship models with performances comparable to many of the existing TIs (Faulon *et al.* 2003; Chemmanattuvalappil *et al.* 2010). In fact, Faulon *et al.* (2003) showed that since both methods are based on graph theory, then one method can be derived from the other. For example, the general relationship between a TI and its signature has been expressed as a dot product between the vector of the occurrence number of atomic signature of height h and the vector of TI values computed for each root of those atomic signatures (Chemmanattuvalappil *et al.* 2010):

$$TI(G) = k {}^h\alpha_G \cdot TI(\text{root}({}^h\Sigma)) \quad (\text{A4.25})$$

Where k is a constant, ${}^h\alpha_G$ is the vector of the occurrence number of atomic signature of height h and $TI(\text{root}({}^h\Sigma))$ is the vector of TI values calculated for each root of atomic signature. It has

been shown that different topological indices of different signature heights can be simultaneously combined using molecular signature and parameterized in an easy to use group contribution format (Chemangattuvalappil *et al.* 2009; Chemangattuvalappil *et al.* 2010).

$$L_i = TI\left(\text{root}\left({}^h\sum\right)\right) \quad (\text{A4.26})$$

$$TI = \sum_{i=1}^N {}^h\alpha_i L_i \quad (\text{A4.27})$$

Remembering that TI can be expressed as chemical and physical properties through QSPR and QSAR quantitative models, then a GCM molecular operator can be formulated:

$$\psi(P) = \sum_{i=1}^N x_i L_i \quad (\text{A4.28})$$

Implementing molecular signatures in this manner requires the use of additional rules for the formation of complete structures. The rules help to ensure that the signatures selected based on the property constraints will connect to form a complete graph without any free bonds. These rules include the handshaking lemma, coloring function (valency of each carbon atom at each level), and the handshaking dilemma, which are discussed in detail by Chemangattuvalappil *et al.* (2010).

Appendix A5

Excipient Case Study Supporting Information

This section contains supporting documentation of the acetaminophen excipient tablet design. Due to the size of the data sets used in the case study, they cannot be displayed in the dissertation in a conventional manner. As a result they have been converted to .mat files and can be provided upon request by contacting Dr. Mario Eden at edenmar@auburn.edu. A list of the data sets that were used is given below:

- $P_{n\text{xp}}$, a set of the pure component IR/NIR properties of the training set
- $R_{n\text{xp}}$, a set of the standardized IR/NIR properties of the training set
- $L_{m\text{xp}}$, a set of the PCA loadings of the training set
- $P_{t\text{xp}}$, a set of the projected IR/NIR properties of the design targets
- $R_{t\text{xp}}$, a set of the standardized IR/NIR properties of the design targets

Appendix A6

IR & NIR Molecular Group Absorbance Frequencies

The IR absorbance frequencies and magnitudes of the functional groups presented below were found in Socrates (2001). The NIR frequencies were found in Workman and Weyer (2008). Only those absorbances verified in these texts were included below. No conjugated double bonds were included, although these can be added later to differentiate between isomers. Also, any intensities listed as variable, were not included.

A6.1. Methine CH

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
Bending (δ)	1360-1320 (1340)	w
Stretching (ν)	2890-2880 (2885)	w
1 st Overtone Bending (2δ)	4049-4029 (4039)	<i>m-w (in NIR)</i>
Combination ($\nu_s + \delta$)	4272-4252 (4262)	<i>m (in NIR)</i>
Combination ($\nu_a + \delta$)	4333-4313 (4323)	<i>m-s (in NIR)</i>

A6.2. Methylene CH₂

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
Scissoring Bend (δ_s)	1480-1440 (1460)	<i>m</i>
Symmetrical Stretching (ν_s)	2870-2840 (2855)	<i>m</i>
Asymmetrical Stretching (ν_a)	2940-2915 (2925)	<i>m-s</i>
Combination ($\nu_a + \delta_s$)	4262-4257 (4260)	<i>m-s (in NIR)</i>
Combination ($\nu_s + \delta_s$)	4336-4332 (4334)	<i>m-s (in NIR)</i>
1 st Overtone Asym. Str. ($2\nu_a$)	5681-5661 (5671)	<i>w (in NIR)</i>
1 st Overtone Sym. Str. ($2\nu_s$)	5690-5670 (5680)	<i>w (in NIR)</i>
Combination ($\nu_a + \nu_s$)	5900-5700 (5800)	<i>m-w (in NIR)</i>
1 st Overtone Comb ($2\nu + \delta$)	7186-7080 (7183)	<i>m-w (in NIR)</i>

2 nd Overtone Sym. Str. ($3\nu_s$)	8257-8237 (8247)	<i>m-w, sh (in NIR)</i>
2 nd Overtone Asym. Str. ($3\nu_a$)	8399-8379 (8389)	<i>m-w, sh (in NIR)</i>

Note Coupling interactions and corrections will need to be added.

A6.3. Hydroxyl OH (w/o H-bond)

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
Bending (δ)	710-570 (640)	<i>m, br</i>
Stretching (ν)	3670-3580 (3625)	<i>m-w, sh</i>
Combination ($\nu + \delta$)	4550-5550 (5050)	<i>m-w, br (in NIR)</i>
1 st Overtone Str. (2ν)	7160-7020 (7090)	<i>m, br (in NIR)</i>
2 nd Overtone Str. (3ν)	10410-10390 (10400)	<i>w (in NIR)</i>

A6.4. Hydroxyl OH (w/ H-bond)

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
Bending (δ)	710-570 (640)	<i>m, br</i>
Stretching (ν)	3550-3230 (3440)	<i>m-s, br</i>
Combination (<i>many types</i>)	4550-5550 (5050)	<i>m-w, br (in NIR)</i>
1 st Overtone Str. (2ν)	6850-6240 (6545)	<i>m, br (in NIR)</i>
1 st Overtone Str. (2ν)	7200-7000 (7100)	<i>m (in NIR)</i>
2 nd Overtone Str. (3ν)	9560-9540 (9550)	<i>w (in NIR)</i>

A6.5. Secondary Alcohol CHOH (w/ H-bond)

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-O OoP. Def. Bend (δ)	390-330 (360)	<i>m-w</i>
C-O IP. Def. Bend (δ_d)	500-440 (470)	<i>W</i>
O-H OoP. Def. Bending (δ_d)	660-600 (630)	<i>m-w, br</i>
C-CO Stretch (ν)	900-800 (850)	<i>m-w</i>
C-O Stretch (ν)	1150-1075 (1113)	<i>S</i>
C-H Def. Bending (δ_d)	1350-1290 (1320)	<i>W</i>
C-H Wag Bend (δ_w)	1400-1330 (1365)	<i>W</i>
O-H + C-H ₂ Coup. Bend. (δ_c)	1430-1370 (1400)	<i>m-s, br</i>
O-H Def. Bend.	1440-1260 (1350)	<i>m-s, br</i>
C-H Stretching (ν)	2890-2880 (2885)	<i>W</i>
O-H Stretching (ν)	3550-3230 (3440)	<i>m-s, br</i>
C-H 1 st Over. Bend. (2δ)	4049-4029 (4039)	<i>m-w (in NIR)</i>
C-H Combination ($\nu_s + \delta$)	4272-4252 (4262)	<i>m (in NIR)</i>
C-H Combination ($\nu_a + \delta$)	4333-4313 (4323)	<i>m-s (in NIR)</i>
O-H Combination (<i>many</i>)	4550-5050 (4800)	<i>m-w, br (in NIR)</i>
O-H Comb. ($\nu + \delta$)	4970-4950 (4960)	<i>w, sh (in NIR)</i>

O-H + C-H Comb. ($\nu + \nu$)	4720-4700 (4710)	w, sh (in NIR)
C-H + OH Combo. ($\delta + \nu$)	6000-5080 (5090)	w, sh (in NIR)
O-H 1 st Overtone Str. (2ν)	6850-6240 (6545)	m, d (in NIR)
O-H 1 st Overtone Str. (2ν)	7150-7000 (7075)	m, d (in NIR)
O-H + 1 st C-H Comb. ($\nu + 2\delta$)	9396-9376 (9386)	m-w (in NIR)
O-H + 2 nd C-O Comb. ($\nu + 3\nu$)	9730-9710 (9720)	w (in NIR)
O-H 2 nd Overtone Str. (3ν)	9560-9540 (9550)	w (in NIR)

A6.6 Primary Alcohol CH₂OH (w/ H-bond)

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-O Def. Bend (δ)	555-395 (475)	m-w
C-O IP. Def. Bend (δ_d)	500-440 (470)	w
O-H OoP. Def. Bending (δ_d)	710-570 (640)	m-w, br
C-CO Stretch (ν)	900-800 (850)	m
C-H ₂ Twist Bend (δ_t)	960-800 (880)	m-w
C-C-O Stretch (ν)	1090-1000 (1045)	S
C-H ₂ Twist. Bending (δ_t)	1300-1280 (1320)	m-w
C-H ₂ Wag Bend (δ_w)	1390-1280 (1335)	m-w
O-H Def. Bend.	1440-1260 (1350)	m-s, br
C-H ₂ Def Bend (δ_d)	1480-1410 (1445)	m-w
C-H ₂ Sym. Stretch (ν_s)	2935-2840 (2888)	m-w
C-H ₂ Asym. Stretch (ν_a)	2990-2900 (2945)	m-w
O-H Stretching (ν)	3550-3230 (3440)	m-s, br
C-H 1 st Over. Bend. (2δ)	4049-4029 (4039)	m-w (in NIR)
C-H Combination ($\nu_s + \delta$)	4272-4252 (4262)	m (in NIR)
C-H Combination ($\nu_a + \delta$)	4333-4313 (4323)	m-s (in NIR)
O-H Combination (many)	4550-5050 (4800)	m-w, br (in NIR)
O-H Comb. ($\nu + \delta$)	4970-4950 (4960)	w, sh (in NIR)
O-H + C-H Comb. ($\nu + \nu$)	4720-4700 (4710)	w, sh (in NIR)
C-H + OH Combo. ($\delta + \nu$)	6000-5080 (5090)	w, sh (in NIR)
O-H 1 st Overtone Str. (2ν)	6850-6240 (6545)	m, d (in NIR)
O-H 1 st Overtone Str. (2ν)	7200-7000 (7100)	m, d (in NIR)
O-H + 1 st C-H Comb. ($\nu + 2\delta$)	9396-9376 (9386)	m-w (in NIR)
O-H + 2 nd C-O Comb. ($\nu + 3\nu$)	9730-9710 (9720)	w (in NIR)
O-H 2 nd Overtone Str. (3ν)	9560-9540 (9550)	w (in NIR)

A6.7. Alcohol CHCH₂OH (w/ H-bond)

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-O Def. Bend (δ)	555-395 (475)	m-w
C-O IP. Def. Bend (δ_d)	500-440 (470)	w
O-H OoP. Def. Bending (δ_d)	710-570 (640)	m-w, br

C-CO Stretch (ν)	900-800 (850)	<i>m</i>
C-H ₂ Twist Bend (δ_t)	960-800 (865)	<i>m-w</i>
C-C-O Stretch (ν)	1045-1025 (1035)	<i>s</i>
C-H ₂ Twist. Bending (δ_t)	1300-1280 (1320)	<i>m-w</i>
C-H ₂ Wag Bend (δ_w)	1390-1280 (1335)	<i>m-w</i>
O-H Def. Bend.	1440-1260 (1350)	<i>m-s, br</i>
C-H ₂ Def Bend (δ_d)	1480-1410 (1445)	<i>m-w</i>
C-H ₂ Sym. Stretch (ν_s)	2935-2840 (2888)	<i>m-w</i>
C-H ₂ Asym. Stretch (ν_a)	2990-2900 (2945)	<i>m-w</i>
O-H Stretching (ν)	3550-3230 (3440)	<i>m-s, br</i>
C-H 1 st Over. Bend. (2δ)	4049-4029 (4039)	<i>m-w (in NIR)</i>
C-H Combination ($\nu_s + \delta$)	4272-4252 (4262)	<i>m (in NIR)</i>
C-H Combination ($\nu_a + \delta$)	4333-4313 (4323)	<i>m-s (in NIR)</i>
O-H Combination (<i>many</i>)	4550-5050 (4800)	<i>m-w, br (in NIR)</i>
O-H Comb. ($\nu + \delta$)	4970-4950 (4960)	<i>w, sh (in NIR)</i>
O-H + C-H Comb. ($\nu + \nu$)	4720-4700 (4710)	<i>w, sh (in NIR)</i>
C-H + OH Combo. ($\delta + \nu$)	6000-5080 (5090)	<i>w, sh (in NIR)</i>
O-H 1 st Overtone Str. (2ν)	6850-6240 (6545)	<i>m, d (in NIR)</i>
O-H 1 st Overtone Str. (2ν)	7200-7000 (7100)	<i>m, d (in NIR)</i>
O-H + 1 st C-H Comb. ($\nu + 2\delta$)	9396-9376 (9386)	<i>m-w (in NIR)</i>
O-H + 2 nd C-O Comb. ($\nu + 3\nu$)	9730-9710 (9720)	<i>w (in NIR)</i>
O-H 2 nd Overtone Str. (3ν)	9560-9540 (9550)	<i>w (in NIR)</i>

A6.8. Aliphatic Ether O

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
Bending (δ_s)	440-420 (430)	<i>w</i>
Symmetrical Stretching (ν_s)	1140-820 (980)	<i>w</i>
Asymmetrical Stretching (ν_a)	1150-1060 (1105)	<i>s</i>

Note Coupling interactions and corrections will need to be added.

A6.9. Ether CH₂O

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-O-C Def. Bending (δ_d)	440-420 (430)	<i>w</i>
C-O-C Asym. Stretching (ν_s)	1140-1085 (1120)	<i>s</i>
C-O-C Sym. Stretching (ν_s)	1140-820 (980)	<i>w</i>
C-H ₂ Def. Bending (δ_d)	1400-1360 (1380)	<i>m</i>
C-H ₂ Wag. Bending (δ_w)	1475-1445 (1460)	<i>m</i>
C-H ₂ Sym. Stretching (ν_s)	2880-2835 (2858)	<i>m</i>
C-H ₂ Asym. Stretching (ν_a)	2955-2920 (2938)	<i>m</i>
C-H + C-O-C Comb. ($\nu_s + \nu_s$)	4080-3980 (4000)	<i>s (in NIR)</i>
C-H ₂ Comb. ($\nu_a + \delta_s$)	4262-4257 (4260)	<i>m-s (in NIR)</i>

C-H ₂ Comb. ($\nu_s + \delta_s$)	4336-4332 (4334)	<i>m-s (in NIR)</i>
C-H ₂ 1 st Ov. Asym. Str. ($2\nu_a$)	5681-5661 (5671)	<i>w (in NIR)</i>
C-H ₂ 1 st Ov. Sym. Str. ($2\nu_s$)	5690-5670 (5680)	<i>w (in NIR)</i>
C-H ₂ Comb. ($\nu_a + \nu_s$)	5900-5700 (5800)	<i>m-w (in NIR)</i>
1 st Ov. Comb. ($2\nu + \delta$)	7186 – 7080 (7183)	<i>m-w (in NIR)</i>
C-H ₂ 2 nd Ov. Sym. Str. ($3\nu_s$)	8257-8237 (8247)	<i>m-w, sh (in NIR)</i>
C-H ₂ 2 nd Ov. Asym. Str. ($3\nu_a$)	8399-8379 (8389)	<i>m-w, sh (in NIR)</i>

A6.10. Aldehyde HCO (w/o H-bond)

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-H Def. Bending (δ_d)	975-780 (877.5)	<i>m-w</i>
C-H IP Rock. Bend. (δ_r)	1375-1350 (1383)	<i>m-s</i>
C-H Stretch Fermi ($\nu + 2\nu$)	1400-1380 (1390)	<i>m-w, d</i>
C=O Stretch (ν)	1790-1710 (1750)	<i>s</i>
C-H Stretch (ν)	2745-2650 (2720)	<i>m-w</i>
C-H Stretch (ν)	2900-2800 (2850)	<i>m-w</i>
C=O + 2 nd C-H Comb. ($\nu + 2\delta$)	4524-4504 (4514)	<i>m-w (in NIR)</i>
C-H + C=O Combo. ($\delta + \nu$)	4758-4738 (4748)	<i>w (in NIR)</i>
C=O + C-H Comb. ($\nu + \nu$)	4760-4445 (4603)	<i>m-w (in NIR)</i>
2 nd C=O + 2 nd C-H Com. ($\nu + 2\delta$)	4898-4878 (4888)	<i>w (in NIR)</i>
2 nd C=O + C-H Com. ($2\nu + \nu$)	7860-7840 (7850)	<i>w (in NIR)</i>

A6.11. Aldehyde HCO (w/ H-bond)

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-H Def. Bending (δ_d)	975-780 (877.5)	<i>m-w</i>
C-H IP Rock. Bend. (δ_r)	1375-1350 (1383)	<i>m-s</i>
C-H Stretch Fermi ($\nu + 2\nu$)	1400-1380 (1390)	<i>m-w, d</i>
C=O Stretch (ν)	1770-1690 (1730)	<i>s</i>
C-H Stretch (ν)	2745-2650 (2720)	<i>m-w</i>
C-H Stretch (ν)	2900-2800 (2850)	<i>m-w</i>
C=O + 2 nd C-H Comb. ($\nu + 2\delta$)	4504-4484 (4494)	<i>m-w (in NIR)</i>
C-H + C=O Combo. ($\delta + \nu$)	4738-4718 (4728)	<i>w (in NIR)</i>
C=O + C-H Comb. ($\nu + \nu$)	4740-4425 (4583)	<i>m-w (in NIR)</i>
C=O + 2 nd C-H Com. ($\nu + 2\delta$)	4878-4858 (4868)	<i>w (in NIR)</i>
1 st C=O + C-H Com. ($2\nu + \nu$)	7860-7840 (7850)	<i>w (in NIR)</i>

A6.12. Saturated Aliphatic Aldehyde CH₂HCO (w/o H-bond)

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-CO IP Def. Bend. (δ_d)	566-520 (543)	<i>m-s</i>

C-C-CO IP Def. Bend. (δ_r)	695-635 (665)	<i>m-s</i>
C-H ₂ Twist. Bending (δ_t)	1300-1280 (1320)	<i>m-w</i>
C-H ₂ Wag Bend (δ_w)	1390-1280 (1335)	<i>m-w</i>
C-H IP Rock. Bend. (δ_r)	1440-1325 (57.5)	<i>m-s</i>
C=O Stretch (ν)	1740-1720 (1730)	<i>s</i>
C-H 1 st Ov. IP Def. Bend. (δ)	2740-2700 (2720)	<i>m-w</i>
C-H Stretch (ν)	2870-2800 (2835)	<i>m-w</i>
C-H Comb. ($\nu_a + \delta_s$)	4262-4257 (4260)	<i>m-s (in NIR)</i>
C-H Comb. ($\nu_s + \delta_s$)	4336-4332 (4334)	<i>m-s (in NIR)</i>
C=O + 2 nd C-H Comb. ($\nu + 2\delta$)	4524-4504 (4514)	<i>m-w (in NIR)</i>
C-H + C=O Combo. ($\delta + \nu$)	4758-4738 (4748)	<i>w (in NIR)</i>
C=O + C-H Comb. ($\nu + \nu$)	4760-4445 (4603)	<i>m-w (in NIR)</i>
2 nd C=O + 2 nd C-H Com. ($2\nu + 2\delta$)	4898-4878 (4888)	<i>w (in NIR)</i>
C-H 1 st Over.Asym. Str. ($2\nu_a$)	5681-5661 (5671)	<i>w (in NIR)</i>
C-H 1 st Over. Sym. Str. ($2\nu_s$)	5690-5670 (5680)	<i>w (in NIR)</i>
C-H Comb. ($\nu_a + \nu_s$)	5900-5700 (5800)	<i>m-w (in NIR)</i>
C-H 1 st Over. Comb. ($2\nu + \delta$)	7186-7080 (7183)	<i>m-w (in NIR)</i>
2 nd C=O + C-H Com. ($2\nu + \nu$)	7860-7840 (7850)	<i>w (in NIR)</i>
C-H 2 nd Over. Sym. Str. ($3\nu_s$)	8257-8237 (8247)	<i>m-w, sh (in NIR)</i>
C-H 2 nd Over. Asym. Str. ($3\nu_a$)	8399-8379 (8389)	<i>m-w, sh (in NIR)</i>

A6.13. Saturated Aliphatic Aldehyde CH₂HCO (w/ H-bond)

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-CO IP Def. Bend. (δ_d)	566-520 (543)	<i>m-s</i>
C-C-CO IP Def. Bend. (δ_r)	695-635 (665)	<i>m-s</i>
C-H ₂ Twist. Bending (δ_t)	1300-1280 (1320)	<i>m-w</i>
C-H ₂ Wag Bend (δ_w)	1390-1280 (1335)	<i>m-w</i>
C-H IP Rock. Bend. (δ_r)	1440-1325 (57.5)	<i>m-s</i>
C=O Stretch (ν)	1720-1700 (1710)	<i>s</i>
C-H 1 st Ov. IP Def. Bend. (δ)	2740-2700 (2720)	<i>m-w</i>
C-H Stretch (ν)	2870-2800 (2835)	<i>m-w</i>
C-H Comb. ($\nu_a + \delta_s$)	4262-4257 (4260)	<i>m-s (in NIR)</i>
C-H Comb. ($\nu_s + \delta_s$)	4336-4332 (4334)	<i>m-s (in NIR)</i>
C=O + 2 nd C-H Comb. ($\nu + 2\delta$)	4504-4484 (4494)	<i>m-w (in NIR)</i>
C-H + C=O Combo. ($\delta + \nu$)	4738-4718 (4728)	<i>w (in NIR)</i>
C=O + C-H Comb. ($\nu + \nu$)	4740-4425 (4583)	<i>m-w (in NIR)</i>
2 nd C=O + 2 nd C-H Com. ($\nu + 2\delta$)	4878-4858 (4868)	<i>w (in NIR)</i>
C-H 1 st Over.Asym. Str. ($2\nu_a$)	5681-5661 (5671)	<i>w (in NIR)</i>
C-H 1 st Over. Sym. Str. ($2\nu_s$)	5690-5670 (5680)	<i>w (in NIR)</i>
C-H Comb. ($\nu_a + \nu_s$)	5900-5700 (5800)	<i>m-w (in NIR)</i>
C-H 1 st Over. Comb. ($2\nu + \delta$)	7186-7080 (7183)	<i>m-w (in NIR)</i>
2 nd C=O + C-H Com. ($2\nu + \nu$)	7840-7820 (7830)	<i>w (in NIR)</i>

C-H 2 nd Over. Sym. Str. (3 ν_s)	8257-8237 (8247)	<i>m-w, sh (in NIR)</i>
C-H 2 nd Over. Asym. Str. (3 ν_a)	8399-8379 (8389)	<i>m-w, sh (in NIR)</i>

A6.14. Vinyl CH₂CH

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C=C Tors. Bend (δ_T)	485-410 (448)	<i>m-s</i>
C=C Eth. Twist. Bend. (δ_t)	600-380 (490)	<i>m-s</i>
C=C Eth. Twist. Bend. (δ_t)	720-410 (565)	<i>w</i>
C-H ₂ OoP Rock. Bend. (δ_r)	980-810 (895)	<i>s</i>
C-H OoP Bending. (δ_r)	1010-940 (975)	<i>s</i>
C-H IP Def. Bend. (δ_d)	1180-1010 (1095)	<i>m-w</i>
C-H ₂ Def. Bend. (δ_d)	1330-1240 (1285)	<i>m</i>
C-H ₂ Sci. Bend. (δ_s)	1440-1360 (1400)	<i>m</i>
C=C Stretching (ν)	1645-1640 (1643)	<i>m-w</i>
C-H ₂ 1 st Overtone Bend (2 δ)	1840-1820 (1830)	<i>w or i</i>
C-H 1 st Overtone Bend (2 δ)	1990-1970 (1980)	<i>w or i</i>
C-H ₂ Sym. Stretch (ν_s)	3070-2930 (3000)	<i>M</i>
C-H Stretch (ν)	3110-2980 (3045)	<i>M</i>
C-H ₂ Asym. Stretch (ν_a)	3150-3000 (3075)	<i>M</i>
C-H ₂ 2 nd Ov. Asy. Bend. (2 δ_a)	4492-4472 (4482)	<i>m (in NIR)</i>
C-H ₂ 2 nd Ov. Bend. (2 δ)	4610-4590 (4600)	<i>w (in NIR)</i>
C-H ₂ 2 nd Ov. Sym. Bend. (2 δ_s)	4780-4670 (4725)	<i>m-w (in NIR)</i>
C-H ₂ 1 st Over. Stretch (2 ν)	6010-5990 (6000)	<i>w (in NIR)</i>
C-H Combo. ($\nu+2\nu$)	6130-6110 (6120)	<i>w (in NIR)</i>

A6.15. Vinylidene CH₂C

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C=C Skeletal Stretch (ν)	470-435 (452)	<i>m-w</i>
C=C Skeletal Stretch (ν)	560-530 (545)	<i>s</i>
C=C Eth. Twist. Bend. (δ_t)	715-680 (698)	<i>w</i>
C-H ₂ OoP Rock. Bend. (δ_r)	895-885 (890)	<i>s</i>
C-H ₂ IP Def. Bend. (δ_d)	1320-1290 (1305)	<i>w</i>
C-H ₂ Sci. Def Bend. (δ_s)	1420-1405 (1413)	<i>w</i>
C=C Stretching (ν)	1675-1625 (1650)	<i>m-w</i>
C-H ₂ 1 st Overtone Bend (2 δ)	1800-1750 (1775)	<i>w</i>
C-H ₂ Sym. Stretch (ν_s)	2985-2970 (2978)	<i>m-w</i>
C-H ₂ Asym. Stretch (ν_a)	3095-3075 (3085)	<i>m-w</i>
C-H ₂ 2 nd Ov. Asy. Bend. (2 δ_a)	4492-4472 (4482)	<i>m (in NIR)</i>
C-H ₂ 2 nd Ov. Bend. (2 δ)	4610-4590 (4600)	<i>w (in NIR)</i>
C-H ₂ 2 nd Ov. Sym. Bend. (2 δ_s)	4780-4670 (4725)	<i>m-w (in NIR)</i>
C-H ₂ 1 st Over. Stretch (2 ν)	6010-5990 (6000)	<i>w (in NIR)</i>

C-H Combo. ($\nu+2\nu$)	6140-6120 (6130)	w (in NIR)

A6.16. *cis*-Vinylene CHCH

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-H Tors. Bend (δ_T)	490-320 (405)	<i>m-s</i>
C=C Skeletal Bend (δ_T)	500-460 (480)	<i>s</i>
-C=CH Def. Bend. (δ_d)	590-440 (515)	<i>m-s</i>
C=C Eth. Twist. Bend. (δ_t)	630-570 (600)	<i>s</i>
C-H Wag. Bend. (δ_w)	790-650 (720)	<i>m-s</i>
C-H Wag. Bend. (δ_w)	1000-850 (925)	<i>m-w</i>
C-H Def. Bend. (δ_d)	1295-1185 (1243)	<i>w</i>
C-H Def. Bend. (δ_d)	1425-1355 (1390)	<i>w</i>
C=C Stretching (ν)	1665-1630 (1648)	<i>m</i>
C-H Stretch (ν)	3040-2980 (3010)	<i>m</i>
C-H Stretch (ν)	3090-3010 (3050)	<i>m</i>
C-H + C=C Combo. Str. ($\nu+\nu$)	4683-4663 (4673)	<i>w (in NIR)</i>
C-H + C=C Combo. Str. ($\nu+\nu$)	4610-4590 (4600)	<i>m-w (in NIR)</i>
C-H 1 st Overtone Stretch (2ν)	5973-5953 (5963)	<i>w (in NIR)</i>

A6.17. *trans*-Vinylene CHCH

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-H Tors. Bend (δ_T)	490-320 (405)	<i>m-s</i>
C=C Skeletal Bend (δ_T)	500-480 (490)	<i>s</i>
-C=CH Def. Bend. (δ_d)	590-440 (515)	<i>m-s</i>
C=C Eth. Twist. Bend. (δ_t)	580-515 (548)	<i>m-s</i>
C-H Wag. Bend. (δ_w)	850-750 (800)	<i>m-w</i>
C-H Wag. Bend. (δ_w)	1000-910 (955)	<i>\nu (w or i)</i>
C-H Def. Bend. (δ_d)	1305-1260 (1283)	<i>\nu (w or i)</i>
C-H Def. Bend. (δ_d)	1340-1355 (1390)	<i>\nu (w or i)</i>
C=C Stretching (ν)	1680-1665 (1673)	<i>m-w</i>
C-H Stretch (ν)	3050-3000 (3025)	<i>m</i>
C-H Stretch (ν)	3065-3015 (3030)	<i>m</i>

A6.18. Aliphatic Methyl CH₃

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
Sym Bend (δ_s)	1390-1370 (1380)	<i>m-s</i>
Asym. Bend. (δ_a)	1465-1440 (1453)	<i>m</i>
Symmetrical Stretching (ν_s)	2885-2865 (2875)	<i>m</i>
Asymmetrical Stretching (ν_a)	2975-2950 (2963)	<i>m-s</i>

Comb. Sym. Bend. ($\delta_u + \delta_u + \delta_u$)	4110-3990 (4100)	<i>s (in NIR)</i>
Combination ($\nu + \delta$)	4420-4370 (4395)	<i>m-s (in NIR)</i>
Com. Asym. Bend. ($\delta_u + \delta_u + \delta_u$)	4425-4375 (4400)	<i>m-s (in NIR)</i>
1 st Overtone Sym. Str. ($2\nu_s$)	5901-5849 (5876)	<i>w (in NIR)</i>
1 st Overtone Asym. Str. ($2\nu_a$)	5925-5885 (5905)	<i>w (in NIR)</i>
1 st Overtone Comb ($2\nu + \delta$)	7273-7253 (7263)	<i>m (in NIR)</i>
1 st Overtone Comb ($2\nu + \delta$)	7365-7345 (7355)	<i>m (in NIR)</i>
2 nd Overtone Stretch (3ν)	8257-8237 (8247)	<i>s (in NIR)</i>

Note The bending combination ($\delta_u + \delta_u + \delta_u$) may actually be a 2nd Overtone (3δ), but that would violate the power hierarchy rule since no 1st order bending mode was identified.

A6.19. Aryl Methyl CH_3

Band	Wavelength Region (cm^{-1})	Relative Intensity & Shape
Rocking Bend. (δ_r)	1060-900 (965)	<i>m-w, d</i>
Rocking Bend. (δ_r)	1130-1000 (1075)	<i>m-w, d</i>
Sym Bend (δ_s)	1405-1355 (1380)	<i>m-s</i>
Asym. Bend. (δ_a)	1470-1400 (1435)	<i>m, d</i>
Asym. Bend. (δ_a)	1480-1430 (1455)	<i>m, d</i>
1 st Over. Asym. Bend. ($2\delta_a$)	2830-2740 (2785)	<i>m-w</i>
1 st Over. Sym. Bend. ($2\delta_a$)	2870-2860 (2865)	<i>m-w</i>
Symmetrical Stretching (ν_s)	2930-2920 (2925)	<i>m-s, d</i>
Asymmetrical Stretching (ν_a)	2955-2935 (2963)	<i>m-s</i>
Symmetrical Stretching (ν_s)	3000-2965 (2983)	<i>m-s, d</i>
Comb. Sym. Bend. ($\delta_u + \delta_u + \delta_u$)	4110-3990 (4100)	<i>s (in NIR)</i>
Combination ($\nu + \delta$)	4420-4370 (4395)	<i>m-s (in NIR)</i>
Com. Asym. Bend. ($\delta_u + \delta_u + \delta_u$)	4425-4375 (4400)	<i>m-s (in NIR)</i>
Com. Str. 2 nd O Bend ($\nu_a + 2\delta_a$)	5660-5640 (5650)	<i>w (in NIR)</i>
1 st Overtone Sym. Str. ($2\nu_s$)	5745-5725 (5735)	<i>w (in NIR)</i>
1 st Overtone Asym. Str. ($2\nu_a$)	5925-5885 (5790)	<i>w (in NIR)</i>
1 st Overtone Comb ($2\nu + \delta$)	7273-7253 (7263)	<i>m (in NIR)</i>
1 st Overtone Comb ($2\nu + \delta$)	7365-7345 (7355)	<i>m (in NIR)</i>
2 nd Overtone Stretch (3ν)	8257-8237 (8247)	<i>s (in NIR)</i>

Note An unsaturated methyl group was used for the bending vibrations since no aryl specific modes were identified by Socrates.

A6.20. Aliphatic Methoxy OCH_3

Band	Wavelength Region (cm^{-1})	Relative Intensity & Shape
C-O Def. Bend. (δ_d)	580-340 (460)	<i>m-w</i>
CH_3/CO Rocking Bend (δ_d)	1190-1100 (1145)	<i>m-w</i>
CH_3 Rock Bend	1235-1155 (1195)	<i>m-w</i>
CH_3 Sym Bend (δ_s)	1460-1420 (1440)	<i>M</i>

CH ₃ Asym. Bend. (δ_a)	1475-1435 (1455)	<i>m, d</i>
CH ₃ Asym. Bend. (δ_a)	1485-1445 (1465)	<i>m, d</i>
C-H ₃ Sym. Str. (ν_s)	2880-2815 (2848)	<i>m, sh</i>
C-H ₃ Asym. Str. (ν_a)	2985-2920 (2950)	<i>m, d</i>
C-H Asym. Str. (ν_a)	3030-2950 (2990)	<i>m, d</i>
CH ₃ Co. S. Bend. ($\delta_u + \delta_u + \delta_u$)	4110-3990 (4100)	<i>s (in NIR)</i>
CH ₃ Combination ($\nu + \delta$)	4420-4370 (4395)	<i>m-s (in NIR)</i>
CH ₃ Co. A. Bend. ($\delta_u + \delta_u + \delta_u$)	4425-4375 (4400)	<i>m-s (in NIR)</i>
OCH ₃ 1 st Ov. Sym. Str. ($2\nu_s$)	5798-5748 (5773)	<i>w (in NIR)</i>
OCH ₃ 1 st Ov. Asym. Str. ($2\nu_a$)	5905-5855 (5880)	<i>w (in NIR)</i>
OCH ₃ 1 st Ov. Comb ($2\nu + \delta$)	7173-7153 (7163)	<i>m (in NIR)</i>
OCH ₃ 1 st Ov. Comb ($2\nu + \delta$)	7345-7325 (7335)	<i>m (in NIR)</i>
CH ₃ 2 nd Ov. Stretch (3ν)	8257-8237 (8247)	<i>m (in NIR)</i>

Note A combination of groups listed in the Chapter 2 and Chapter 7 where ranges were extended to include both ranges.

A6.21. Aryl Methoxy OCH₃

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-O Def. Bend. (δ_d)	580-340 (460)	<i>m-w</i>
C-O Bend (δ)	1050-1010 (1030)	<i>m</i>
CH ₃ /CO Rocking Bend (δ_r)	1190-1100 (1145)	<i>m-w</i>
CH ₃ Rock Bend (δ_r)	1235-1155 (1195)	<i>m-w</i>
C-O Bend (δ)	1310-1210 (1260)	<i>m</i>
CH ₃ Sym Bend (δ_s)	1460-1420 (1440)	<i>m</i>
CH ₃ Asym. Bend. (δ_a)	1475-1435 (1455)	<i>m, d</i>
CH ₃ Asym. Bend. (δ_a)	1485-1445 (1465)	<i>m, d</i>
C-H ₃ Sym. Str. (ν_s)	2860-2815 (2850)	<i>m, sh</i>
C-H ₃ Asym. Str. (ν_a)	2985-2920 (2950)	<i>m-w, d</i>
C-H Asym. Str. (ν_a)	3005-2965 (2985)	<i>m-w, d</i>
CH ₃ Co. S. Bend. ($\delta_u + \delta_u + \delta_u$)	4110-3990 (4100)	<i>s (in NIR)</i>
CH ₃ Combination ($\nu + \delta$)	4420-4370 (4395)	<i>m-s (in NIR)</i>
CH ₃ Co. A. Bend. ($\delta_u + \delta_u + \delta_u$)	4425-4375 (4400)	<i>m-s (in NIR)</i>
OCH ₃ 1 st Ov. Sym. Str. ($2\nu_s$)	5798-5748 (5773)	<i>w (in NIR)</i>
OCH ₃ 1 st Ov. Asym. Str. ($2\nu_a$)	5905-5855 (5880)	<i>w (in NIR)</i>
OCH ₃ 1 st Ov. Comb ($2\nu + \delta$)	7173-7153 (7163)	<i>m (in NIR)</i>
OCH ₃ 1 st Ov. Comb ($2\nu + \delta$)	7345-7325 (7335)	<i>m (in NIR)</i>
CH ₃ 2 nd Ov. Stretch (3ν)	8257-8237 (8247)	<i>m (in NIR)</i>

Note A combination of groups listed in the Chapter 2 and Chapter 7 where ranges were extended to include both ranges.

A6.22. Aryl OH (w/o H-bond)

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-OH IP Bending (δ)	450-375 (413)	w
C-O Stretch (ν)	1260-1150 (1205)	s
O-H IP Bending (δ)	1410-1300 (1355)	s
O-H Stretching (ν)	3620-3590 (3605)	m, sh
O-H Comb. Str. ($\nu + \nu$)	4883-4683 (4783)	m-s (in NIR)
O-H + C-O Comb. Str. ($\nu + \nu$)	4967-4767 (4867)	m-w (in NIR)
O-H Combination ($\nu + \delta$)	5049-4849 (4949)	w (in NIR)
1 st Overtone Str. (2ν)	7140-6940 (7040)	m-w, br (in NIR)
2 nd Overtone Str. (3ν)	10010-9990 (10000)	w (in NIR)

A6.23. Aryl OH (w/ H-bond)

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-OH IP Bending (δ)	450-375 (413)	w
O-H OoP. Def. Bending (δ_d)	720-600 (660)	s, br
C-O Stretch (ν)	1260-1180 (1220)	s
O-H IP Bending (δ)	1410-1310 (1360)	s
O-H Stretching (ν)	3250-3000 (3125)	m, br
O-H Comb. Str. ($\nu + \nu$)	4883-4683 (4783)	m-s (in NIR)
O-H + C-O Comb. Str. ($\nu + \nu$)	4967-4767 (4867)	m-w (in NIR)
O-H Combination ($\nu + \delta$)	5049-4849 (4949)	w (in NIR)
O-H 1 st Overtone Str. (2ν)	6470-6270 (6370)	w, br (in NIR)
O-H 1 st Overtone Str. (2ν)	7000-6700 (6850)	w, br (in NIR)
2 nd Overtone Str. (3ν)	9920-9900 (9910)	w, d (in NIR)
2 nd Overtone Str. (3ν)	10300-10280 (10290)	w, d (in NIR)

A6.24. Tetramethyl C(CH₃)₃

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-C Skeletal Bend (δ_s)	930-925 (928)	m
C-C Skeletal Bend (δ_s)	1010-990 (1000)	m-w
C-C Skeletal Bend (δ_s)	1225-1165 (1195)	m
C-C Skeletal Bend (δ_s)	1255-1245 (1250)	m
C-CH ₃ Sym. Bend. (δ_s)	1395-1350 (1365)	m-s
C-CH ₃ Sym. Bend. (δ_s)	1420-1375 (1398)	m
C-CH ₃ Asym. Bend. (δ_a)	1475-1435 (1455)	m
C-H Sym. Stretching (ν_s)	2885-2865 (2875)	m
C-H Asym. Stretching (ν_a)	2975-2950 (2963)	m-s
C-H Co. S. Bend. ($\delta_u + \delta_{u+} + \delta_u$)	4110-3990 (4100)	s (in NIR)

C-H Combination ($\nu + \delta$)	4420-4370 (4395)	<i>m-s (in NIR)</i>
Com. Asym. Bend. ($\delta_u + \delta_{u+} \delta_u$)	4425-4375 (4400)	<i>m-s (in NIR)</i>
1 st Overtone Sym. Str. ($2\nu_s$)	5901-5849 (5876)	<i>w (in NIR)</i>
1 st Overtone Asym. Str. ($2\nu_a$)	5925-5885 (5905)	<i>w (in NIR)</i>
1 st Overtone Comb ($2\nu + \delta$)	7273-7253 (7263)	<i>m (in NIR)</i>
1 st Overtone Comb ($2\nu + \delta$)	7365-7345 (7355)	<i>m (in NIR)</i>
2 nd Overtone Stretch (3ν)	8257-8237 (8247)	<i>s (in NIR)</i>

A6.25. Alkyl Peroxide OO

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
O-O Stretch (ν)	900-800 (850)	<i>w</i>
C-O Stretch (ν)	1150-1030 (1090)	<i>m-s</i>

A6.26. Saturated Aliphatic Ester COO

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-O-C Sym. Stretch (ν_s)	1160-1050 (1105)	<i>s</i>
C-O-C Asym. Stretch (ν_a)	1275-1185 (1230)	<i>s</i>
C=O Stretch (ν)	1750-1725 (1738)	<i>s</i>
C=O 1 st Overtone ($2\nu_s$)	3460-3440 (3450)	<i>w</i>

A6.27. Saturated Aliphatic Methyl Ester COOCH₃

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
Unlisted	450-430 (440)	<i>m-s</i>
CO-O Rocking Bend (δ_r)	530-340 (435)	<i>w</i>
C-C-O Sym. Stretch (ν_s)	1160-1050 (1105)	<i>s</i>
C-O Stretch (ν)	1175-1155 (1165)	<i>s</i>
C-C-O Asym. Stretch (ν_a)	1275-1185 (1230)	<i>s</i>
O-CH ₃ Stretch (ν)	1315-1195 (1245)	<i>s</i>
Unlisted	1370-1350 (1360)	<i>w</i>
CH ₃ Sym. Def. Bend (δ)	1460-1420 (1440)	<i>m-w</i>
CH ₃ Asym. Def. Bend (δ)	1465-1420 (1443)	<i>m-s</i>
CH ₃ Asym. Def. Bend (δ)	1485-1435 (1455)	<i>m</i>
C=O Stretch (ν)	1750-1725 (1738)	<i>s</i>
CH ₃ Sym. Stretch (ν)	3000-2860 (2930)	<i>m</i>
CH ₃ Asym. Stretch (ν)	3030-2950 (2990)	<i>m-w</i>
CH ₃ Asym. Stretch (ν)	3050-2980 (3015)	<i>m-w</i>
C=O 1 st Overtone ($2\nu_s$)	3460-3440 (3450)	<i>w</i>
CH ₃ Com. S. Bend. ($\delta_u + \delta_{u+} \delta_u$)	4110-3990 (4100)	<i>s (in NIR)</i>
CH ₃ Comb. ($\nu + \delta$)	4420-4370 (4395)	<i>m-s (in NIR)</i>
CH ₃ Com. A. Bend. $\delta_u + \delta_{u+} \delta_u$	4425-4375 (4400)	<i>m-s (in NIR)</i>

Com. Str. 2 nd O Bend ($\nu_a + 2\delta_a$)	5660-5640 (5650)	w (in NIR)
OCH ₃ 1 st Ov. Sym. Str. ($2\nu_s$)	5901-5725 (5803)	w (in NIR)
OCH ₃ 1 st Ov. Asym. Str. ($2\nu_a$)	5925-5855 (5890)	w (in NIR)
OCH ₃ 1 st Ov. Comb ($2\nu + \delta$)	7273-7153 (7253)	m (in NIR)
OCH ₃ 1 st Ov. Comb ($2\nu + \delta$)	7365-7325 (7345)	m (in NIR)
CH ₃ 2 nd Ov. Stretch (3ν)	8257-8237 (8247)	m-s (in NIR)

Note The unlisted absorptions could not be cross-referenced within Socrates, so they were left as undefined. Also, no NIR data on ester groups was available in Workman and Weyer (2008), so a combination of aldehydes and methoxys were used to approximate the behavior.

A6.28. Saturated Aliphatic Ethyl Ester CH₃CH₂COO

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C-O-C Def Bend (δ)	370-250 (310)	m-w
C-O-C Def Bend (δ)	395-305 (350)	m-w
CO-O Rocking Bend (δ_r)	485-365 (425)	m-w
CO OoP Rocking Bend (δ_r)	700-550 (625)	w
CH ₂ Rocking Bend (δ_r)	825-775 (800)	w
C-C str (ν)	940-850 (895)	w
CH ₃ Rock. Bend (δ_r)	1150-1080 (1115)	w
C-C-O Sym. Stretch (ν_s)	1160-1050 (1105)	s
CH ₃ Rock. Bend (δ_r)	1195-1135 (1165)	w
C-C-O Asym. Stretch (ν_a)	1275-1185 (1230)	s
CH ₂ Twist. Bend (δ_T)	1340-1325 (1333)	m-w
CH ₂ Wag. Bend (δ_w)	1385-1335 (1360)	m-w
CH ₃ Sym. Def. Bend (δ)	1390-1360 (1375)	m-s
CH ₃ Asym. Def. Bend (δ)	1480-1435 (1458)	m
OCH ₂ Def. Bend. (δ)	1490-1460 (1475)	m-w
C=O Stretch (ν)	1750-1725 (1738)	s
CH ₃ Stretch (ν)	2920-2860 (2890)	w
CH ₃ Sym. Stretch (ν_s)	2930-2890 (2910)	w
CH ₃ Asym. Stretch (ν_a)	2995-2930 (2963)	m
C=O 1 st Overtone ($2\nu_s$)	3460-3440 (3450)	w
CH ₃ Com. S. Bend. ($\delta_u + \delta_u + \delta_u$)	4110-3990 (4100)	s (in NIR)
C-H Comb. ($\nu_a + \delta_s$)	4262-4257 (4260)	m-s (in NIR)
C-H Comb. ($\nu_s + \delta_s$)	4336-4332 (4334)	m-s (in NIR)
CH ₃ Comb. ($\nu + \delta$)	4420-4370 (4395)	m-s (in NIR)
CH ₃ Com. A. Bend. $\delta_u + \delta_u + \delta_u$	4425-4375 (4400)	m-s (in NIR)
Com. Str. 2 nd O Bend ($\nu_a + 2\delta_a$)	5660-5640 (5650)	w (in NIR)
CH ₃ 1 st Ov. Sym. Str. ($2\nu_s$)	5901-5725 (5803)	w (in NIR)
CH ₃ 1 st Ov. Asym. Str. ($2\nu_a$)	5925-5855 (5890)	w (in NIR)
CH ₃ 1 st Ov. Comb ($2\nu + \delta$)	7273-7153 (7253)	m (in NIR)

CH ₃ 1 st Ov. Comb (2ν+δ)	7365-7325 (7345)	<i>m (in NIR)</i>
CH ₃ 2 nd Ov. Stretch (3ν)	8257-8237 (8247)	<i>m-s (in NIR)</i>

Note The unlisted absorptions could not be cross-referenced within Socrates, so they were left as undefined. Also, no NIR data on ester groups was available in Workman and Weyer, so a combination of aldehydes and methoxys were used to approximate the behavior.

A6.29. Acrylate Ester CH₂CHCOO

Band	Wavelength Region (cm ⁻¹)	Relative Intensity & Shape
C=C Tors. Bend (δ _T)	485-410 (448)	<i>m-s</i>
CO-O Rocking Bend (δ _r)	485-365 (425)	<i>m-w</i>
C=C Eth. Twist. Bend. (δ _t)	600-380 (490)	<i>m-s</i>
C-O-C Def Bend (δ)	675-660 (668)	<i>m</i>
CO OoP Rocking Bend (δ _r)	700-550 (625)	<i>w</i>
=CH ₂ Twist Bend (δ _T)	810-800 (805)	<i>m-s</i>
CH ₂ Rocking Bend (δ _r)	825-775 (800)	<i>w</i>
C-C str (ν)	940-850 (895)	<i>w</i>
=CH ₂ Wag. Bend (δ _w)	970-960 (965)	<i>s</i>
C-H Def. Wag (δ _w)	990-980 (985)	<i>m</i>
C-H OoP Bending. (δ _r)	1010-940 (975)	<i>s</i>
C-C Skel. Bend (δ)	1070-1065 (1068)	<i>m</i>
CH ₃ Rock. Bend (δ _r)	1150-1080 (1115)	<i>w</i>
C-C-O Sym. Stretch (ν _s)	1160-1050 (1105)	<i>s</i>
C-H IP Def. Bend. (δ _d)	1180-1010 (1095)	<i>m-w</i>
unlisted	1200-1195 (1198)	<i>s</i>
C-C-O Asym. Stretch (ν _a)	1275-1185 (1230)	<i>s</i>
=CH Rock. Bend (δ _r)	1290-1270 (1280)	<i>m</i>
unlisted	1290-1280 (1285)	<i>s</i>
=CH ₂ Def Bend (δ)	1420-1400 (1410)	<i>m</i>
C-H ₂ Sci. Bend. (δ _s)	1440-1360 (1400)	<i>m</i>
C=C Stretch (ν)	1635-1615 (1625)	<i>m</i>
C=C Stretch (ν)	1650-1630 (1640)	<i>m-s</i>
C=O Stretch (ν)	1725-1710 (1718)	<i>s</i>
C-H ₂ 1 st Overtone Bend (2δ)	1840-1820 (1830)	<i>w or i</i>
C-H 1 st Overtone Bend (2δ)	1990-1970 (1980)	<i>w or i</i>
C-H ₂ Sym. Stretch (ν _s)	3070-2930 (3000)	<i>m</i>
C-H Stretch (ν)	3110-2980 (3045)	<i>m</i>
C-H ₂ Asym. Stretch (ν _a)	3150-3000 (3075)	<i>m</i>
C=O 1 st Overtone (2ν _s)	3460-3440 (3450)	<i>w</i>
C-H ₂ 2 nd Ov. Asy. Bend. (2δ _a)	4492-4472 (4482)	<i>m (in NIR)</i>
C-H ₂ 2 nd Ov. Bend. (2δ)	4610-4590 (4600)	<i>w (in NIR)</i>
C-H ₂ 2 nd Ov. Sym. Bend. (2δ _s)	4780-4670 (4725)	<i>m-w (in NIR)</i>
C-H ₂ 1 st Over. Stretch (2ν)	6010-5990 (6000)	<i>w (in NIR)</i>

C-H Combo. ($\nu+2\nu$)	6130-6110 (6120)	w (in NIR)

A6.30. Methacrylate Ester $\text{CH}_2\text{C}(\text{CH}_3)\text{COO}$

Band	Wavelength Region (cm^{-1})	Relative Intensity & Shape
C=C Skeletal Stretch (ν)	470-435 (452)	<i>m-w</i>
C=C Skeletal Stretch (ν)	560-530 (545)	<i>s</i>
C-O-C Def Bend (δ)	660-645 (653)	<i>m</i>
C=C Eth. Twist. Bend. (δ_t)	715-680 (698)	<i>w</i>
C-C Skel Bend (δ)	825-805 (815)	<i>m-s</i>
C-H ₂ OoP Rock. Bend. (δ_r)	895-885 (890)	<i>s</i>
=CH ₂ Wag. Bend (δ_w)	950-935 (943)	<i>s</i>
C-C Skel. Bend (δ)	1010-990 (1000)	<i>m</i>
C-C Skel. Bend (δ)	1020-1000 (1010)	<i>m</i>
C-O-C Sym. Stretch (ν_s)	1160-1150 (1155)	<i>s</i>
C-O-C Asym. Stretch (ν_a)	1275-1185 (1230)	<i>s</i>
unlisted	1310-1290 (1300)	<i>s</i>
C-H ₂ IP Def. Bend. (δ_d)	1320-1290 (1305)	<i>w</i>
=CH Rock. Bend (δ_r)	1335-1315 (1325)	<i>m</i>
CH ₃ Sym Bend (δ_s)	1390-1370 (1380)	<i>m-s</i>
=CH ₂ Def Bend (δ)	1420-1400 (1410)	<i>m</i>
CH ₃ Asym. Bend. (δ_a)	1465-1440 (1453)	<i>m</i>
C=C Stretch (ν)	1650-1630 (1640)	<i>m</i>
C=O Stretch (ν)	1725-1710 (1718)	<i>s</i>
C-H ₂ 1 st Overtone Bend (2δ)	1800-1750 (1775)	<i>w</i>
CH ₃ Sym. Stretching (ν_s)	2885-2865 (2875)	<i>m</i>
C-H ₂ Sym. Stretch (ν_s)	2985-2970 (2978)	<i>m-w</i>
CH ₃ Asym. Stretching (ν_a)	2975-2950 (2963)	<i>m-s</i>
C-H ₂ Asym. Stretch (ν_a)	3095-3075 (3085)	<i>m-w</i>
C=O 1 st Overtone ($2\nu_s$)	3460-3440 (3450)	<i>w</i>
CH ₃ Com. S. Bend. ($\delta_u + \delta_{u+} \delta_u$)	4110-3990 (4100)	<i>s (in NIR)</i>
CH ₃ Combination ($\nu + \delta$)	4420-4370 (4395)	<i>m-s (in NIR)</i>
CH ₃ Com. A. Bend. ($\delta_u + \delta_{u+} \delta_u$)	4425-4375 (4400)	<i>m-s (in NIR)</i>
C-H ₂ 2 nd Ov. Asy. Bend. ($2\delta_a$)	4492-4472 (4482)	<i>m (in NIR)</i>
C-H ₂ 2 nd Ov. Bend. (2δ)	4610-4590 (4600)	<i>w (in NIR)</i>
C-H ₂ 2 nd Ov. Sym. Bend. ($2\delta_s$)	4780-4670 (4725)	<i>m-w (in NIR)</i>
CH ₃ 1 st Ov. Sym. Str. ($2\nu_s$)	5901-5849 (5876)	<i>w (in NIR)</i>
CH ₃ 1 st Ov. Asym. Str. ($2\nu_a$)	5925-5885 (5905)	<i>w (in NIR)</i>
C-H ₂ 1 st Over. Stretch (2ν)	6010-5990 (6000)	<i>w (in NIR)</i>
C-H Combo. ($\nu+2\nu$)	6140-6120 (6130)	<i>w (in NIR)</i>
CH ₃ 1 st Ov. Comb ($2\nu + \delta$)	7273-7253 (7263)	<i>m (in NIR)</i>
CH ₃ 1 st Ov. Comb ($2\nu + \delta$)	7365-7345 (7355)	<i>m (in NIR)</i>
CH ₃ 2 nd Ov. Stretch (3ν)	8257-8237 (8247)	<i>s (in NIR)</i>

A6.31. Other Groups

Additional groups available in the library, but not included in this dissertation are α -Pyranose and β -Pyranose.