

**A Systematic Property Based Approach for Molecular Synthesis Using Higher
Order Molecular Groups and Molecular Descriptors**

by

Nishanth Gopalakrishnan Chemmangattuvalappil

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 13, 2010

Key Words: Molecular Design, Topological Indices, Molecular Signatures

Copyright 2010 by Nishanth Gopalakrishnan Chemmangattuvalappil

Approved by

Mario R. Eden, Associate Professor of Chemical Engineering, Auburn University, Chair
Christopher B. Roberts, Professor of Chemical Engineering, Auburn University
Ram B. Gupta, Professor of Chemical Engineering, Auburn University
Jin Wang, Assistant Professor of Chemical Engineering, Auburn University
Mahmoud M. El-Halwagi, Professor of Chemical Engineering, Texas A&M University

Abstract

In this work, algorithms have been developed for the design of molecules corresponding to the optimum performance of a process. The concept of property clustering has been extended into molecular design based on second and third order group contribution methods. An algebraic approach has been developed utilizing higher order molecular groups built from first order groups. The significant aspect of the aforementioned method is that both the application range and reliability of the molecular property clustering technique are considerably increased by incorporating second and third order estimation. A methodology has been developed for incorporating the property contribution predicted using combined group contribution and connectivity indices into the design framework in case the property contributions of any of the molecular groups of interest are not available in literature. For the design of simple mono-functional molecules, a modified visual approach has been used whereas for the design of more complicated structures and/or for treating more than three properties at a time, an algebraic method has been developed.

Until now, most reverse property prediction algorithms are based on group contribution methods. However, a variety of properties can be predicted using Quantitative Structure Activity/Property Relationships (QSAR/QSPR) models. QSAR models make use of topological indices to predict physical properties and biological activities. In this dissertation, a new algorithm has been developed to include topological

index based property models into the reverse problem formulation framework. This algorithm makes use of the concept of molecular signature descriptors to incorporate a variety of different topological indices on a common platform. A large number of environmental, safety and health related constraints can be now investigated as a part of the integrated process and molecular design. An algorithm for the enumeration of the molecular structures has been developed with very low degeneracy. In the last part, a general framework has been proposed to simultaneously integrate process and product design problems with flowsheet design. This methodology will identify the best candidate molecules that provide the optimum process performance with minimum energy utilization. The dissertation concludes with a list of potential areas where more study can be conducted based on the developed algorithms.

Acknowledgments

I would like to express my profound gratitude to Dr. Mario R. Eden for his constant support, encouragement and assistance. Special recognition is given for his guidance and direction. He has always been a great source of information and inspiration. I would like to thank my research committee members, Dr. Mahmoud El-Halwagi at Texas A&M University, Dr. Ram Gupta, Dr. Christopher Roberts and Dr. Jin Wang for their valuable comments and suggestions. My sincere thanks to my collaborators Charles Solvason, Dr. Fadwa Eljack and Susilpa Bommareddy for their brilliant ideas and feedback. Thanks are also due to my friends and co-workers, Dr. Norman Sammons, Dr. Jeffrey Seay, Wei Yuan, Subin Hada and Gregory Vaughan at Auburn University. Special gratitude and appreciation is given to my parents, Gopalakrishnan C.C. and Vijayakumari T.V., my wife Suchithra, and my sister Dhanya for all their support and encouragement. Finally, I would like to thank the faculty and staff in the Department of Chemical Engineering at Auburn University for making my graduate research experience at Auburn a memorable and rewarding one.

Table of Contents

Abstract.....	ii
Acknowledgments.....	iv
List of Figures.....	ix
List of Tables	xi
Nomenclature.....	xiv
1. Introduction.....	1
2. Theoretical Background.....	9
2.1. Chemical Product Design	9
2.2. Mathematical Formulation of Chemical Product Design ..	14
2.3. Types of Properties and Estimation Techniques.....	17
2.4. Mixture Design	19
2.4.1. Design of Experiments.....	19
2.4.2. Mixture Design of Experiments.....	21
2.5. Group Contribution Methods.....	23
2.5.1. Initial Efforts.....	23
2.5.2. Group Contribution Models with Higher Levels	25
2.6. Topological Indices and Property Prediction.....	29
2.6.1. Connectivity Indices	32

2.6.2. Edge Adjacency Index	37
2.6.3. Shape Indices	39
2.6.4. Wiener Indices	40
2.6.5. The Hoyosa Topological Index.....	40
2.7. Connectivity Indices and GC ⁺ Method	41
2.8. Molecular Signature Descriptors	43
2.8.1. Current Status In Inverse Design	43
2.8.2. Development of Molecular Signature.....	49
2.8.3. Application of Signature Descriptors in Property Prediction	55
2.9. Flowsheet Property Model.....	56
2.10. Summary	59
3. Basics of Computer Aided Molecular Design	62
3.1. Computer Aided Molecular Design Framework.....	62
3.2. Computer Aided Molecular Design Techniques.....	66
3.3. Property Models.....	67
3.4. Reverse Problem Formulation	69
3.5. Reverse Problem Formulation Methodology.....	70
3.6. Summary	71
4. Integrated Process and Molecular Design.....	73
4.1. Property Clustering Techniques.....	73
4.1.1. Property Operator and Cluster Formulation	73
4.1.2. Conservation Rules	76
4.1.3. Visualization Techniques.....	78

4.1.4. Identification of Feasibility Region	80
4.2. Molecular Property Operators and Clusters.....	82
4.3. Visual Solution of Molecular Design Problem.....	85
4.4. Limitations of the Visual Approach in Molecular Design.....	90
4.5. An Algebraic Approach for Molecular Synthesis with Higher Order Groups	92
4.5.1. General Problem Statement	95
4.5.2. Algebraic Approach for Solving the Molecular Design Problem.....	95
4.5.3. Algebraic Molecular Design Algorithm	102
4.6. Introduction of GC+ Models into the Cluster Space	104
4.6.1. GC ⁺ Algorithm for Visual Solution of a Molecular Design problem.....	106
4.6.2. GC ⁺ Algorithm for Algebraic Solution of a Molecular Design Problem ..	109
4.7. Molecular Signatures in Reverse Problem Formulations	110
4.7.1. First Order Connectivity Index	111
4.7.2. Kier-Hall Shape Index of Order 1.....	112
4.7.3. Reverse Problem Formulation using Molecular Signatures	114
4.7.4. Signature Based Algorithm for Molecular Design	115
4.7.5. Expression of Group Contribution Models with Signatures.....	126
4.7.6. Property Models with Different Signature Heights	129
4.7.7. Enumeration of Molecular Structures from Signatures	131
4.7.8. Stepwise Procedure for Solving a Molecular Design Problem.....	136
4.8. General Framework for Integrated Flowsheet and Molecular Design.....	137
4.9. Summary	139
5. Case Studies	141

5.1. Design of Blanket Wash Solvent	141
5.2. Metal Degreasing Solvent Design	151
5.2.1. Visual Solution.....	154
5.2.2. Algebraic Solution	158
5.3. Design of Alkyl Substituent for the Fungicide DD	162
5.3.1. Problem Statement	162
5.3.2. Solution of Design Problem with Two Types of Property Models	163
5.3.3. Solution of Design Problem with Different Topological Indices	173
5.4. Acid Gas Removal	175
5.4.1. Problem Statement	175
5.4.2. Process Design	177
5.4.3. Molecular Design.....	178
5.4.4. Proof of Concept for Integrated Flowsheet and Molecular Design	187
6. Conclusions and Future Work	191
6.1. Major Achievements	191
6.2. Future Work	195
6.2.1. Statistical Tools for Product Design	195
6.2.2. Integrated Process and Product design	196
6.2.3. Inclusion of More Sophisticated Descriptors.....	198
6.2.4. Exploration of Biochemical Reaction Pathways.....	199
6.2.5. Design of Reactions	201
References.....	203
Appendix: Group Contribution and Connectivity Indices Data	218

List of Figures

Figure 2.1 General Chemical Product Design	12
Figure 2.2 Product Design Steps	13
Figure 2.3 Property Estimation Models	19
Figure 2.4 Response Surface of the Second Order Model.....	21
Figure 2.5 Mixture Design Plots.....	22
Figure 2.6 Multilevel Approach for Property Estimation using GC Method	26
Figure 2.7 Example of Hydrogen Suppressed Graphs.....	31
Figure 2.8 Molecular Skeleton of 3,3 Dimethyl Pentane.....	34
Figure 2.9 Atomic Signatures up to Height 3	52
Figure 2.10 Molecular Signature Tree	54
Figure 3.1 Iterative Molecular Design	62
Figure 3.2 Multilevel Approach for Product Design	65
Figure 3.3 Simultaneous Consideration of Process and Product Design.....	70
Figure 3.4 Reverse Problem Formulation.....	71
Figure 4.1 Visualization of Intra-stream Conservation of Clusters	77
Figure 4.2 Visualization of Inter-stream Conservation of Clusters	77
Figure 4.3 Mixing of Streams	79
Figure 4.4 Feasibility Region on a Ternary Diagram	82
Figure 4.5 Mixing of Molecular Groups.....	88

Figure 4.6 Simultaneous Process and Product Design Framework	90
Figure 4.7 Second Order Group Formation	94
Figure 4.8 Mixing of CI Group with GC Group.....	108
Figure 4.9 Coloring of Atomic Signature	120
Figure 4.10 Illustration of Connectivity Principles.....	123
Figure 4.11 Integrated Process-Product-Flowsheet Design Framework	138
Figure 4.12 Flowchart of the Integrated Process-Product-Flowsheet Design	139
Figure 5.1 Cluster Diagram for Degreaser Design	157
Figure 5.2 Visual Solution of Molecular Design Problem	158
Figure 5.3 Acid Gas Removal Flowsheet	176
Figure 5.4 Best Five Solutions to Acid Gas Removal Problem.....	186

List of Tables

Table 2.1 Types of Products	10
Table 2.2 Group Contribution Models	27
Table 2.3 Adjustable Parameters in Group Contribution Models	28
Table 2.4 Values of K_{C-X} Parameters	39
Table 4.1 Visual Molecular Design Algorithm	85
Table 4.2 Algebraic Approach Algorithm	103
Table 4.3 GC ⁺ Model Algorithm	109
Table 4.4 CI Calculation using Signatures	111
Table 4.5 Signature Equivalent of Topological Indices	113
Table 4.6 Consistency of Signatures	124
Table 5.1 Property Constraints for Blanket Wash Solvent	143
Table 5.2 Property Operators and Reference Values	143
Table 5.3 Adjustable Parameters	143
Table 5.4 Normalized Molecular Property Operator Values	144
Table 5.5 Property Data of Selected Molecular Fragments	144
Table 5.6 Second Order Groups and Their Contributions	147
Table 5.7 Valid Formulations and Their Properties	147
Table 5.8 Groups for Ring Compounds	148
Table 5.9 Second Order Groups and Their Contributions for Cyclic Structures.....	149

Table 5.10 Valid Cyclic Compounds and Their Properties	150
Table 5.11 Property Constraints for the Degreaser Problem	152
Table 5.12 Property Operators and Reference Values for Degreaser Design.....	154
Table 5.13 Normalized Molecular Property Operator Values	154
Table 5.14 Atom and Bond Indices	155
Table 5.15 First Order Connectivity Indices	156
Table 5.16 CI Property Contributions of CI Groups	156
Table 5.17 Possible Higher Order Groups and Their Property Contributions	159
Table 5.18 Final Solution Set for Degreaser Design	160
Table 5.19 Upper and Lower Bounds for Fungicide Properties	163
Table 5.20 Signatures of Height Two for Alkanes	166
Table 5.21 Solution in Terms of Signatures	171
Table 5.22 Possible Alkyl Substituents	172
Table 5.23 New Solution for Alkyl Substituent	175
Table 5.24 Property and Flowrate Data for Acid Gas Removal Problem	177
Table 5.25 Property Targets for Molecular design	178
Table 5.26 Property Operators and Targets	181
Table 5.27 New Property Targets for Acid Gas Absorbent Design	187
Table 5.28 Driving Force Data	189
Table 5.29 Energy Index Values	189
Table A.1 First Order Group Contribution Property Data.....	219
Table A.2 Second Order Group Contribution Property Data	223
Table A.3 Third Order Group Contribution Property Data	226

Table A.4 First Order GC Data for Acentric Factors and Liquid Molar Volume	228
Table A.5 Second Order GC Data for Acentric Factors and Liquid Molar Volume	230
Table A.6 Regressed Parameters for Different Atom Types in the CI Method.....	232

Nomenclature

AUP	Augmented property index
C_j	Property cluster for property j
D	Number of degrees
deg	Degree
FBN	Free bond number
FBN_g	Free bond number associated with group g
GCM	Group contribution method
$G(x)$	Molecular subgraph of atom x
h	Height of signature
H_{fus}	Heat of fusion
h_{fus0}	Adjustment parameter used in the estimation of heat of fusion
h_{fus1}	Contribution of first order group for the estimation of heat of fusion
h_{fus2}	Contribution of second order group for the estimation of heat of fusion
ΔH_v	Standard heat of vaporization at 298 K
h_{v0}	Adjustment parameter used in the estimation of heat of vaporization
h_{v1}	First order group contribution for estimation of heat of vaporization
h_{v2}	Second order group contribution for estimation of heat of vaporization

M	Number of edges
N_g	Total number of first order molecular groups
N_s	Total number of second order molecular groups
N_r	Number of rings in molecular structure
n_f	Number of functional groups in the main chain
n_g	Molecular group
n_{gr}	Groups forming the ring
n_o	Number of carbon atoms in the main chain
P_{jg}	Contribution to property j from group g
R	Number of rings
T_b	Boiling point
t_{bo}	Adjustment parameter used in the estimation of boiling point
t_{b1}	Contribution of first order group for estimation of boiling point
t_{b2}	Contribution of second order group for estimation of boiling point
TI	Topological index
V	Vertex
T_m	Melting point
t_{mo}	Adjustment parameter used in the estimation of melting point
t_{m1}	Contribution of first order group for estimation of melting point
t_{m2}	Contribution of second order group for estimation of melting point
x_s	Fractional contribution

Greek symbols

$\psi_j (P_j)$	Molecular property operator of the j^{th} property
----------------	---

$\psi_j^{ref}(P_{ji})$	Reference operator for j^{th} property
Ω_j	Normalized property operator for property j
η	Number of occurrences of one first order group in a second order group
σ	Molecular signature
α	Number of each signature
θ	Property function
κ	Shape index
χ	Connectivity index

1. Introduction

The selection of products/product mixtures that give the optimum performance of a process is a critical issue for a design engineer. The process performance is usually understood in terms of physical properties and on many occasions, the physical properties of the product rather than their chemical structure determine the suitability of a specific product as the input to the process. For example, in the design of a blanket wash solvent, the primary focus of designer are the solubility parameter, flammability, vapor pressure etc. of the solvent. Molecular design algorithms generally require target properties to design the molecules. At the same time, to identify the target properties that give the optimum process performance, the process parameters are to be considered as well. Therefore, for obtaining the optimal solution for this type of problem, it is necessary to have a methodology to represent the product performance in terms of measurable physical properties and identify the molecule/mixture that gives the property targets corresponding to the optimum process performance.

In spite of the relationship between the process design and product design problems, they have been traditionally considered as two separate problems because the product design part is generally considered as outside the scope of chemical engineering design. Therefore, the product identified by chemists (without considering the process design aspects) may not provide the best solution corresponding to the optimum process performance and this makes the process of identifying the suitable molecule/mixture an iterative process. Therefore, on most occasions, the chemists try to design the product

based more on expert knowledge, trial and error, heuristics and experimentation based on intuition rather than any specific scientific reasoning. The attempts to develop a procedure to pass the information between the process and product designer brought about the development of reverse problem formulations (Eden *et al.*, 2004).

The conventional way to treat a combined process-product design problem was through mathematical modeling. This approach considers the molecular design as an optimization problem. The objective in such an optimization problem is to minimize the error between the sought values and the values attained by the current design. The advantage with this approach is in most of the cases, it is possible to represent the problem in terms of known mathematical expressions. The system of equations formed consists of balance equations, constraint expressions and constitutive equations (Russel *et al.*, 2000). The non-linear nature of the constitutive equations, which is often the case with most structure-property relationships, makes it difficult to obtain convergence during the computational stages. In such cases, by following decomposition techniques, the values of a subset of intensive variables are determined that match the required property targets. This can be considered as the reverse of a property estimation problem. The target properties are estimated by solving a reverse simulation problem, where for the given values of design and input variables, the desired range of property values can be obtained (Gani & Pistikopoulos, 2002).

The reverse problem formulation decouples the complicated property models from the system of equations and the conventional forward problem can be divided into two reverse problems. The first reverse problem solves the balance and constraint equations in terms of properties to provide the design targets. The second reverse

problem then solves the constitutive equations to identify the operating conditions/products to match the property targets set from the first reverse problem. Decoupling the constitutive expressions from the system of equations makes the system linear and achieving convergence is easier. In this way, the reverse problem formulation (RPF) lowers the complexity of the problem without compromising the accuracy (Eden *et al.*, 2004).

The RPF provides a property-based platform to link process and product design problems since the process performance can be represented in terms of the properties, and the properties form the input to an molecular design problem. However, such an algorithm can be followed only if there is a way to track the properties. The concept of property clustering provides the necessary tools to track properties. Property clustering techniques have recently been extended into molecular design to develop molecular clusters that allow the molecular groups to be combined to match a set of target properties (Eljack *et al.*, 2007). The property models available in group contribution methods formed the link connecting the molecular structure and properties in the above-mentioned algorithm.

The purpose of this work is to expand the range of applications of the integrated process and product synthesis approach. The product design part of this approach, even though it conceptually bridged the gap between process and product design is limited in terms of its practical applications in its present form. Three major limitations of the present method are being addressed in this dissertation. The first one is the limitations of the property models that can be applied in that algorithm. The algorithm by Eljack *et al.* (2007) is based on first order group contribution methods. This method provides a good

starting tool in several product design problems. Nevertheless, its applicability is limited to the design of simple molecular structures because the first order group contribution method has limited accuracy especially when dealing with polyfunctional molecules and cyclic molecules. In addition, first order groups cannot capture proximity effects or differentiate between isomers. A variety of newer group contribution methods tried to address this issue by considering the effects of the combinations of certain molecular groups in the chemical structure (Constantinou & Gani, 1994; Marrero & Gani, 2001; Conte *et al.*, 2008). These combinations of molecular groups are known as higher order groups. In order to increase the range and applicability of the design, higher levels of group contribution methods also have to be represented in the cluster domain and considered in generating molecular structures. However, unlike first order groups, the second and third order groups are not linearly represented in a molecule. In addition, since the higher order molecular groups are based on the effects due to the combination of the constituent molecular building blocks, their effects cannot be predicted before knowing the complete molecular structure. Therefore, a new algorithm must be developed for their systematic inclusion into the cluster domain.

In the visual approach developed by Eljack *et al.* (2007), the number of property targets has to be three for the simultaneous consideration of process and product requirements. In their approach, different molecular groups are mixed/combined according to a set of rules developed to obtain different candidate solutions. However, the number of properties of interest may be more than three on many occasions. Similarly, the generation of a complete potential candidate set is very significant for product design problems because, as mentioned in section three (figure 3.2), there are further stages in a

product design algorithm after the elimination based on the group contribution based techniques. Therefore, it is very important to make sure that none of the potential candidates are overlooked in the initial stages of product design because the possibility that a molecule not considered could be the optimum candidate based on parameters other than group contribution method based tools cannot be ignored. Therefore, there should be a generalized procedure to generate all potential solutions of a problem to match the given set of property targets.

The motivation to produce more sustainable and environmental friendly chemicals to meet the consumer needs has increased considerably over the last decade (Kokossis & Yang, 2009). Therefore, it is important to have a systematic methodology to design chemicals that possess both the consumer specified attributes and environmentally acceptable characteristics. Most biological and environmental properties are structure dependent and group contribution techniques are not available or reliable for the determination of these properties. However, a lot of work has already been done to categorize atoms or molecules systematically based on their structure and to relate these assignments to their biological activities and properties. These relationships are termed as Quantitative Structure-Activity Relationships (QSAR) and Quantitative Structure-Property Relationships (QSPR) (Kier & Hall, 1986). QSAR and QSPR can provide viable tools for the determination of many properties from molecular structure information. However, most QSAR and QSPR techniques are very property specific and thousands of molecular structural descriptors are now available in literature corresponding to different properties (Randic & Basak, 2001). In spite of this, very few attempts have been made to make use of the available QSAR/QSPR relationships to solve inverse design problems.

This is because, compared to the group contribution models, the QSAR/QSPR relationships have very complicated formulations due to the highly non-linear nature of most of the topological indices used in developing such relationships. In addition, unlike molecular groups, there is no one to one mapping possible from the solution of an optimization problem to the final molecular structures (Visco *et al.*, 2002; Faulon *et al.*, 2003a). The recently introduced concept of molecular signature description (Visco *et al.*, 2002; Faulon *et al.*, 2003b) is a potential tool for the reverse problem formulation techniques explained in chapter 3 of this dissertation. The motivation behind developing algorithms to introduce signature description into the reverse problem formulation techniques is that, it is an already proven fact that many molecular descriptors can be written in terms of their signatures and many such relationships are linear in nature. Therefore, one single algorithm will have the potential to handle different molecular descriptors (Visco *et al.*, 2002; Faulon *et al.*, 2003b). In addition, the enumeration of molecular structures from the solution of an inverse problem is challenging. Even though, a few stochastic techniques are available to solve this problem (Venkatasubramanian *et al.*, 1994; Sheridan & Kearsley, 1995; Venkatasubramanian *et al.*, 1995), there have been very few attempts to solve this problem using a deterministic approach. Therefore, an algorithm that can solve this type of problems using a deterministic approach would widen the applicability of reverse problem formulations to a different domain of problems.

The dissertation has been distributed in six chapters. Chapter 2 covers most of the background information including details of the nature of process design problems, some current product design techniques, the basics of group contribution methods, topological

indices and their applications in QSAR/QSPR expressions and a brief description of molecular signature descriptors and their application to property prediction. The chapter 3 covers the current state of the art in the field of computer aided molecular design, the role of property models, the concept of reverse problem formulations and the integrated process and product design framework. Chapter 4 starts with the basics of property clustering and first order molecular property clusters. Then, it covers the systematic development of second and third order molecular property operators. The systematic procedure for the development of an algebraic algorithm for the application of these operators in the solution of integrated process and product design is also presented. The next section describes the procedure followed for the introduction of connectivity index based models into the clustering framework and the steps followed in the development of an algorithm for their application in solving integrated process and product design. The developed visual and algebraic approaches are described. In the next section, the application of molecular signature descriptors in solving molecular design problems is described. Finally, the general framework to integrate flowsheet design techniques to process and molecular design problems is presented. Chapter 5 provides four application examples for the algorithms developed in chapter 4. The first example is a blanket wash solvent design problem to illustrate the development of higher order molecular groups and the algorithm for their introduction into an algebraic solution framework. The second example illustrates the application of the algorithm developed for the introduction of combined connectivity index/group contribution models into the reverse problem algorithm. The third example is a simple molecular design problem that shows the applications of molecular signature descriptors. The fourth example is an integrated

flowsheet and molecular design problem. The last chapter covers the major achievements and conclusions from this project and highlights some of the future works that can be done based on the techniques developed in this dissertation. Most of the group contribution data and connectivity index data are provided in appendix.

2. Theoretical Background

2.1. Chemical Product Design

Chemical product design is an emerging branch of chemical engineering. In the past, the development of new chemical products has always been left to chemists and the chemical engineering community generally focused on the process design aspects and ignored all product related issues other than purity. Due to this, the product design has always been considered separately from the process design with no feedback between each other. This approach often leads to the generation of sub-optimal solutions. In addition, most chemical products currently in use have been developed after scientific experimentation based on knowledge of existing products that has been largely based on heuristics and expert knowledge. This approach often limits the scope of the output solutions because of the inability to produce non-intuitive solutions. The innumerable options available for such a methodology ultimately make this technique researcher specific and many times based on intuition. Therefore, there is a need for a comprehensive and systematic methodology for solving product design problems.

Cussler has classified all the chemical products into three broad classes (Cussler *et al.*, 2010): commodities, molecular products and performance chemicals. Commodities are bulk chemicals and the focus of chemical engineers while producing these chemicals are traditionally on designing the processes to produce them economically. The second class of chemicals are molecules with specific applications like pharmaceutical products, and the key to market them depends on the speed of the discovery and the ability to

introduce them into the market immediately after the discovery. In the third class of products, the value will be added depends on the specific microstructure. The key to the marketability will be the function and the benefits that they provide like the flavor a chemical provides to the ice cream or the shine a certain polishing material can provide to the shoe. A summary of the types and focus involved in the different classes of products are given in table 2.1 (Cussler *et al.*, 2010):

Table 2.1: Types of Products

	<i>Commodities</i>	<i>Molecules</i>	<i>Performance</i>
Key	Cost	Speed	Function
Basis	Unit operations	Chemistry	Microstructure
Risk	Feedstock	Discovery	Science

Hill (2009) identified the need for a new mindset along with new chemical engineering approaches for the solution of product design problems and termed the emergence of this field as a new paradigm. This is because, the chemical engineering community generally ignored product related issues with the exception of purity. The process related issues were the only areas of interest for chemical engineers. The mindset behind solving the different classes of process design problems follows the same approach. Here, the focus of design will be on obtaining the process with minimum cost. The design of a new product however, requires a different understanding of the profit, which may not be readily converted into a set of mathematical expressions. In order to successfully solve a product design problem, the designer has to identify both the process requirements and product specifications and to systematically generate a finite number of

potential candidate solutions to satisfy the problem requirements. The experimentation can now be limited to the systematically obtained candidates because it is not possible to conduct experiments with all available options.

A chemical product design problem can be stated like this: Identify a chemical product (molecule or mixture) that satisfies a set of desired needs. So, a product design problem can be considered as an inverse property prediction problem where the attributes are represented in terms of physical properties (Gani & O'Connell, 2001).

The purpose of product engineering is not to substitute the traditional experimental techniques and/or heuristics followed to design a molecule. Instead, the product design engineers aim to systematically eliminate the numerous unsuitable options and reduce the search space to a finite set of options. Therefore, the chemical product design can be considered as a phase in the overall product development operation that should precede a well defined experimental program, design and analysis (Hill, 2009).

Cussler and Moggridge (2001) have identified the principal steps involved in a product design process. Specific to each problem, solution strategies are to be developed in each step:

1. Identify customer needs
2. Generate ideas to meet the needs
3. Select among ideas
4. Manufacture product

This framework is a simplified yet generalized representation of what product design is all about. Each step must be defined more elaborately for different specific classes of problems. However, the framework applies to all kinds of problems.

The first step is traditionally considered as a topic handled by marketing experts and chemical engineer's tasks usually start from step 2. However, a recent work has incorporated the first step as a part of an optimization framework as shown in figure 2.1 (Smith & Ierapetritou, 2009). The motivation behind forming such a customer integrated approach for product design is the realization that, the driving force for a product-centered industry is the consumer needs (Stephanopoulos, 2003). Smith and Ierapetritou (2009) developed a product design methodology for which the inputs are the consumer inputs and economic criteria. The design problem has been formulated as a biobjective optimization problem that ensures the consumer influence in design trade-off considerations.

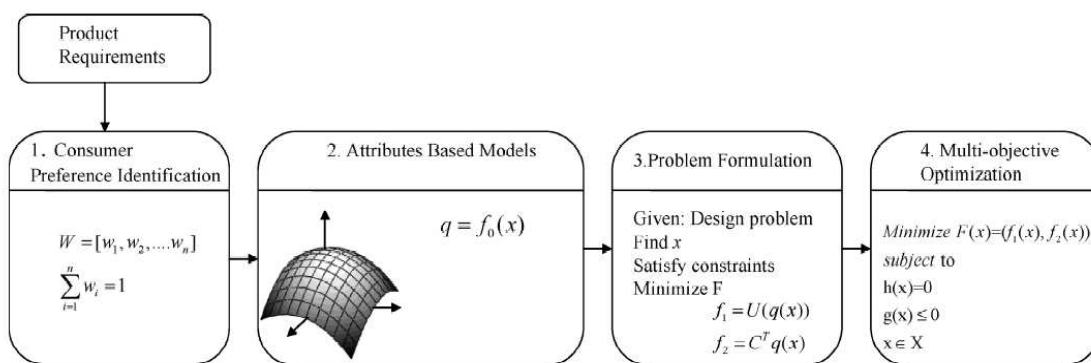


Figure 2.1: General Chemical Product Design (Smith & Ierapetritou, 2009)

In addition, chemical engineers are in a better position to refine the understanding of customer needs because of their ability to analyze what is physically possible (Hill,

2009). Similarly, once the product to be manufactured is finalized, the actual production of it is more specific to the compound and needs to be treated accordingly. The heart of the product design problem is steps 2 and 3 and so, the focus of the work in this area has to be to generate and select among ideas.

Step two has been traditionally performed by chemists according to the specifications provided by process engineers and are typically based on heuristics and expert knowledge. The compounds they identify in this step are analyzed by process design engineers to decide the suitability of the supplied options in step three. If none of the options are practical, their conclusions will be send back to the chemists for more options. Therefore, this is an iterative process as described in figure 2.2.

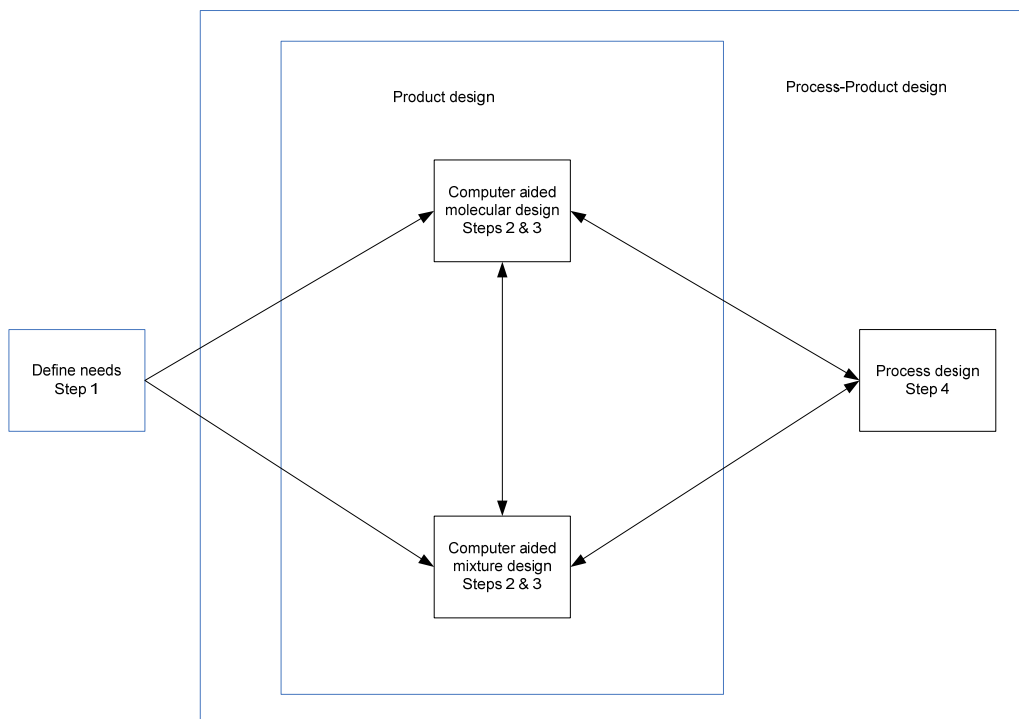


Figure 2.2: Product Design Steps (Gani, 2004a)

The computational complexity involved in identifying a suitable molecule with desirable properties can never be underestimated. Without a systematic approach to track the specific solutions, the number of potential solutions will be prohibitively high even with very restrictive criteria. For instance, if the search space is limited to all alkane molecules up to C_{22} , there will be 38 million different isomers (Davidson, 2002). As the number of atoms under consideration increases, the number of candidates increases and results in combinatorial explosion. For example, even if the search is restricted to acyclic molecules with three double bonds, the number of structural isomers for $C_4H_{10}N_2O_3S_2$ will be 16 million (Contreras *et al.*, 1994)

The traditional approach above points out that, since the process and product design steps are done separately, the product identified may not be corresponding to the optimum process performance since the product design has been done without the knowledge of property targets corresponding to the optimum process performance. A recently developed integrated process and product design approach (Eljack *et al.*, 2007b; Eljack & Eden, 2008) provides an alternate way to solve this problem. This approach and the various tools used will be discussed later in this dissertation.

2.2. Mathematical Formulation of Chemical Product Design

If the focus of the product is on the macroscopic properties, the product design can be considered as a combination of molecular design and mixture design (Achenie *et al.*, 2003). All different types of product design problems can be represented using the following set of generalized mathematical expressions (Gani, 2004a).

$$F_{Obj} = \max\{C^T y + f(x)\} \quad (2.1)$$

$$h_1(x) = 0 \quad (2.2)$$

$$h_2(x) = 0 \quad (2.3)$$

$$h_3(x, y) = 0 \quad (2.4)$$

$$l_1 \leq g_1(x) \leq u_1 \quad (2.5)$$

$$l_2 \leq g_2(x, y) \leq u_2 \quad (2.6)$$

$$l_3 \leq By + Cx \leq u_3 \quad (2.7)$$

In the above expressions, x is the vector of continuous variables like fraction in a mixture, flowrates etc., y is the vector represents the presence or absence of a group, compound, operation, etc., $h_1(x)$ is a set of equality constraints corresponding to process design specifications, $h_2(x)$ is a set of equality constraints corresponding to process model equations, $h_3(x, y)$ is a set of equality constraints related to molecular structure generation, mixing rules for properties, etc, $g_1(x)$ is a set of inequality constraints related to process design specifications, $g_2(x, y)$ is a set of inequality expressions corresponding to specific problems related to the product design and $f(x)$ is the vector of objective functions. For process design problems, $f(x)$ will typically be a non-linear function and for an integrated process and product design problem, $f(x)$ typically represent more than one non-linear expression. B and C represent the matrix containing fixed data in constraint defined by eq. (2.7).

Depending on the specific nature of the problem, some or all of the equations/constraints listed above may be used. A few different types of product design problems are as follows (Gani, 2004a):

- 1) Satisfy only the constraint in eq. (2.6): A product design problem based on a database search. Here, the objective would be to identify from the dataset, the compounds matching the property constraints. Here, the molecular structure generation or the application of property models is not necessary.
- 2) Satisfy constraints in eqs. (2.4) and (2.6): Here, the molecular structures are generated on the basis of property model in eq. (2.4) subject to the constraints in eq. (2.6).
- 3) Satisfy the objective function and constraints in eqs. (2.4) and (2.6): In this problem, the optimum molecular structure has been identified according to the objective function given in eq. (2.1) subjected to the product design constraints in eq. (2.6). The property model given in eq.(2.4) is used to generate the molecular structure. However, there is no guarantee that the solution obtained is an optimal solution because of the non-linear equations that constitute the property models in eq. (2.4).
- 4) Satisfy all the constraints: This type of problems identifies the set of products corresponding to the process requirements. Therefore, this is a simultaneous process-product design problem. This will generate feasible, but not necessarily optimal solutions because of the non-linear nature of the property models in both process and product design specifications.

- 5) Solve all the equations: This is an integrated process-product design problem. The non-linear nature of the objective function and the process model equations will make this a complex mixed integer non-linear programming (MINLP) problem.

In all these kinds of approaches, the properties either need to be provided or predicted through property models (Achenie *et al.*, 2003). Therefore, except for the first type of product design problems, the application range of any developed product design methodology is limited to the accuracy of the involved property models. Therefore, the success of product design methodologies will depend on the included property constraints and the process/product model.

2.3. Types of Properties and Estimation Techniques

Gani and Constantinou (1996) proposed a three tier classification for different properties as primary, secondary and functional.

Primary properties Properties that can be estimated from the molecular structure variables. Examples include normal boiling point, normal melting point, heat of vaporization at 298 K etc.

Secondary properties Pure component properties which are dependent on other properties. Examples are solubility parameter, density at a given temperature, etc.

Functional properties Pure component properties which are dependent on temperature and/or pressure. Examples are density, vapor pressure, enthalpy, etc

Apart from these general classifications, there are a number of high level performance characteristics, which are difficult to estimate. These kinds of properties

involve the taste of food products, aroma of fragrances, various mechanical properties, etc. Since many of these properties are dynamic and the design objectives can only be specified in terms of a time-evolution profile, the modeling process will be very challenging. Different types of hybrid approaches can be applied for solving such problems such as combining molecular modeling with kinetic phenomena to obtain prediction accuracy (Ghosh *et al.*, 2000).

Depending on the types of target properties and expected range of accuracy, different types of property models can be used for product design. Gani and Constantinou (1996) have proposed the classification of property models shown in figure 2.3. However, most of the types of models shown in figure 2.3 are suitable only for forward problems because of the computational complexities involved in the quantum mechanics calculations. The semi-empirical and empirical models also possess many computational difficulties in a traditional mathematical programming based approach. However, different tools have been developed recently to incorporate many of such methods into inverse design formulations (Venkatasubramanian *et al.*, 1995; Camarda & Maranas, 1999; Sahinidis *et al.*, 2003; Papadopoulos & Linke, 2006; Eljack & Eden, 2008; Solvason *et al.*, 2009).

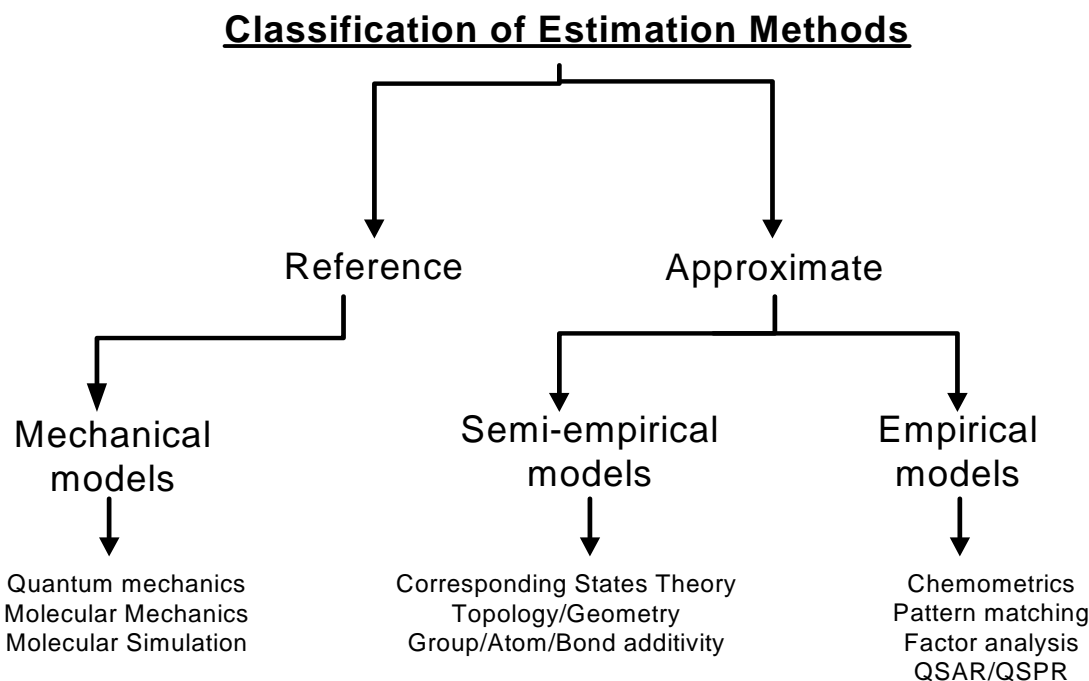


Figure 2.3: Property Estimation Models (Achenie *et al.*, 2003)

2.4. Mixture Design

2.4.1. Design of Experiments

Design of experiments (DOE) is a statistical method to plan and execute experiments in a systematic manner so that maximum information can be gained from the experiments. DOE is a potential tool in the field of product design because, as mentioned previously, the product design in actual practice depends heavily on results obtained through experiments. To develop a model, the first step is to identify the factors that affect the variable of interest. In the next step, a model is postulated to represent the effect of the factors on the response of interest of the variable. The objective is to optimize the response. In the next step, the experimental points are placed to which the model can be fitted. In the final step, the model adequacy is tested. There may be many iterations until the experimenter decides the accuracy is good enough (Cornell, 2002).

The accuracy of this method depends on the adequacy of the model equation and the location of the design points. The polynomial model is the most commonly selected model to represent a response surface since it can be expanded through a Taylor series to improve accuracy (Cornell, 2002). Generally, first or second degree models will be adequate to represent the surface (Montgomery, 2005). Figure 2.4 shows one example of a response surface plot for a second degree model where the response y corresponding to the factors x_1 and x_2 are plotted. The fitted first order and second order equations have the following general form:

$$y = \beta_o + \sum_{i=1}^u \beta_i x_i + \varepsilon \quad (2.8)$$

$$y = \beta_o + \sum_{i=1}^u \beta_i x_i + \sum_{i \leq j}^u \sum_{j \geq i}^u \beta_{ij} x_i x_j + \varepsilon \quad (2.9)$$

Here, y is the response, x_i and x_j are the factors affecting the response. The β values are the regression coefficients and ε is the error observed in the response. Note that, in the second order equation, there is a term corresponding to the interaction effects of the factors involved.

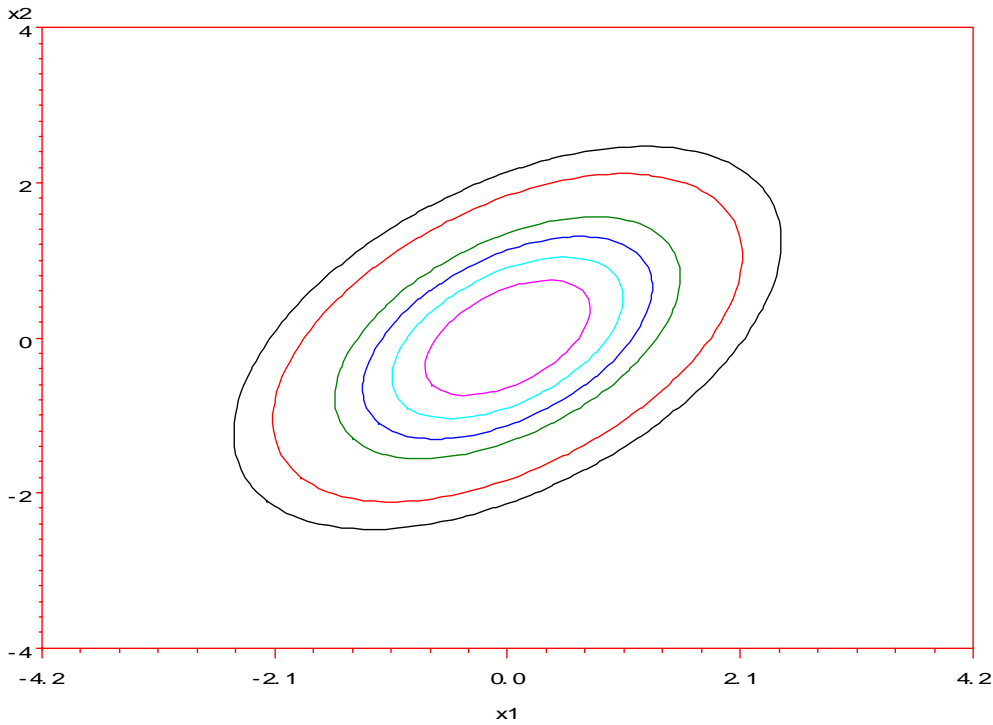


Figure 2.4: Response Surface of the Second Order Model

2.4.2 Mixture Design of Experiments

Mixture design of experiments (MDOE) is an extension of DOE in which the factors are the chemical constituents. Therefore, the constituent fractions will sum to one and every constituent fraction must have a value between zero and one. Figure 2.5 shows one example of a mixture design plot where the density of a mixture made from the components x_1 , x_2 and x_3 is plotted. However, this relationship violates the condition that the factors are independent and random and thereby imposes a colinearity effect. Therefore, even though the model can still be used, it will affect the interpretation of the regression coefficients. However, because of this condition, it is possible to represent the mixture data on a simplex. Scheffe developed the first simplex-lattice designs, which many researchers consider to be the foundation of mixture design (Cornell, 2002).

According to Scheffe models, the response surface can be represented in terms of only the pure component and interaction terms. However, due to the colinearity, the regressors will not provide the true interpretation of the pure component or interaction effects.

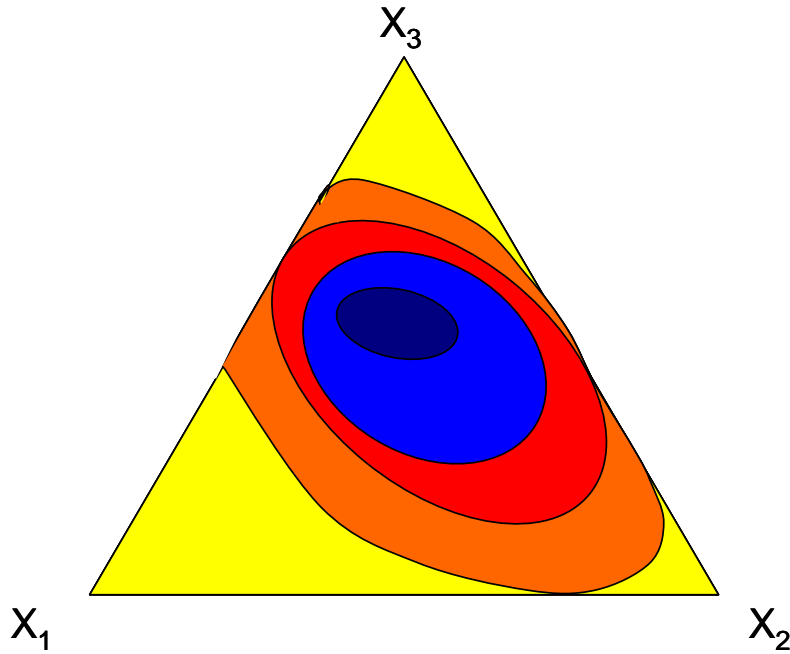


Figure 2.5: Mixture Design Plots

In order to address the limitations of Scheffe models, the Cox model (Cox, 1971) was introduced. Even though the Cox model removes the colinearities introduced by the relation between the mixture constituents, it leaves the secondary colinearities introduced by the constraints in the constituent ranges (Solvason *et al.*, 2008). One solution to this issue may be the use of decomposition techniques like Principal Component Regression (PCR) and Partial Least Squares (PLS) (Kettanah-Wold, 1992). Applying the property clustering technique is another way of treating this issue (Solvason *et al.*, 2008).

2.5. Group Contribution Methods

2.5.1. Initial Efforts

Computer Aided Molecular Design (CAMD) techniques have become a significant part of process and product design because of their ability to predict and design molecules with a given set of properties. In all CAMD algorithms, it is necessary to have a systematic method to evaluate whether the designed structures satisfy the property constraints set by the process from well-defined molecular building blocks. Therefore, almost all CAMD techniques use group contribution methods (GCM) to verify whether the generated molecules exhibit the specified set of desirable properties (Harper *et al.*, 1999). In additive group contribution methods (Benson, 1968; Ambrose, 1978, 1980; Joback & Reid, 1987; Horvath, 1992), a molecule is considered as a collection of various simple groups. The property function of a molecule has been estimated as a summation of the property contributions of all the molecular groups present in the molecular structure.

$$f(X) = \sum_i N_i C_i \quad (2.10)$$

Here, $f(X)$ is a function of the actual property X , C_i is the contribution of the molecular group i that occurs N_i times

These contributions are estimated through regression of large amounts of experimental data. Group contribution methods are indispensable tools for property prediction of molecules from their structures especially when the experimental values for the properties are not available. They are simple and yet provide reasonably accurate

results for many properties. These methods can provide quick estimates of properties without much computational complexity and errors (Constantinou *et al.*, 1993). In the case of simple compounds, GCM can provide accurate trends due to the addition of new functional groups to the existing structure. During molecular synthesis, this will help to generate molecular structures that meet a specific property in a systematic way from basic molecular groups (Joback & Stephanopoulos, 1989).

However, as the complexity of the molecule increases, the accuracy of first order GCM becomes less reliable. They generally cannot capture proximity effects or differentiate between isomers (Kehiaian, 1983; Wu & Sandler, 1989, 1991). The simplified representation of molecular structure in any of the above-mentioned methods ignores many of the concepts in organic chemistry and quantum mechanics like resonance, conjugation and various interactions among groups (Mavrovouniotis, 1990). So, several attempts have been made to make the GCM more general and reliable (Fedors, 1982; Reid *et al.*, 1987; Constantinou *et al.*, 1993).

The ABC method introduced by Constantinou *et al.* (1993) is of particular importance. The ABC method is based on the contributions of atoms and bonds in the properties of different conjugate forms of a molecular structure. Here, the property of a molecule has been estimated as the linear combination of contributions from all the conjugate forms of the molecule. However, the generation and enumeration of the conjugate forms is computationally challenging. Nevertheless, this method provided the basis for future methods, which did not require such computational effort (Constantinou & Gani, 1994).

2.5.2. Group Contribution Models with Higher Levels

A new GC approach was put forward by Constantinou and Gani (1994) in which the property estimation is done in two stages. In this approach, two types of molecular building blocks are defined. The basic level is known as first order groups and the next higher level is called second order groups. The second order groups have first order groups as their building blocks. They essentially represent different types of interactions among the first order groups and the effects of certain molecular group combinations to the property of the final molecule. The second order groups can provide a better description of compounds with many functional groups and differentiate among isomers. However, even the second order groups may not be able to provide a good representation of poly-ring compounds and open-chain polyfunctional compounds with more than four carbon atoms in the main chain (Marrero & Gani, 2001). Therefore, a further level of molecular groups have been identified and their property contributions have been regressed (Marrero & Gani, 2001) for use in group contribution methods. The formation of third order groups is analogous to the second order groups, but the focus is on a different class of molecular groups. The third order groups focus on multi-ring compounds, fused ring compounds and compounds with many functional groups in the structure. Similar to the second order groups, third order groups also have first order groups as their building blocks. The property estimation model developed in this approach has the following form:

$$f(X) = \sum_i N_i C_i + w \sum_j M_j D_j + z \sum_k O_k E_k \quad (2.11)$$

Here, $f(X)$ is a function of the actual property X , C_i is the contribution of first order group i that occurs N_i times, D_j the contribution of second order group j that occurs M_j times and E_k the contribution of third order group k that occurs O_k times in the molecule. The constants w and z can have values zero or unity depending on how many levels of estimation are of interest.

The pictorial representation of the property estimation technique using higher order group contribution techniques has been shown in figure 2.6 (Conte *et al.*, 2008). The properties estimated using this technique and the corresponding property functions are listed in table 2.2. The universal constants for each property function are given in table 2.3. Four properties, for which property functions and group contributions are estimated in two different articles (Constantinou *et al.*, 1995; Conte *et al.*, 2008) up to second order level, are also included in the table 2.2.

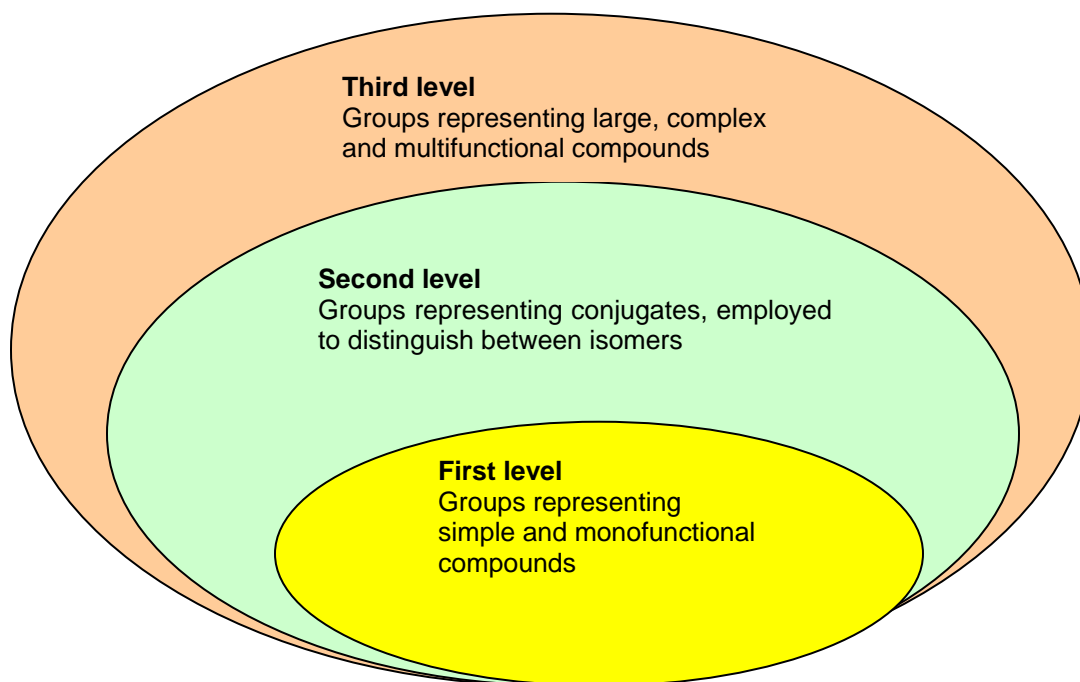


Figure 2.6: Multilevel Approach for Property Estimation using GC Method

Table 2.2: Group Contribution Models

<i>Property</i>	<i>Property Function</i>	<i>Group Contribution Terms</i>
Normal melting point, T_m	$\exp(T_m/T_{m0})$	$\sum_i N_i T_{m1i} + \sum_j M_j T_{m2j} + \sum_k O_k T_{m3k}$
Normal boiling point, T_b	$\exp(T_b/T_{b0})$	$\sum_i N_i T_{b1i} + \sum_j M_j T_{b2j} + \sum_k O_k T_{b3k}$
Critical temperature, T_c	$\exp(T_c/T_{c0})$	$\sum_i N_i T_{c1i} + \sum_j M_j T_{c2j} + \sum_k O_k T_{c3k}$
Viscosity, η	$\ln(\eta)$	$\sum_i N_i \eta_{i1} + \sum_j M_j \eta_{j2}$
Critical volume, V_c	$V_c - V_{c0}$	$\sum_i N_i V_{c1i} + \sum_j M_j V_{c2j} + \sum_k O_k V_{c3k}$
Standard Gibbs energy, G_f	$G_f - G_{f0}$	$\sum_i N_i G_{f1i} + \sum_j M_j G_{f2j} + \sum_k O_k G_{f3k}$
Critical pressure, P_c	$(P_c - P_{c1})^{-0.5} - P_{c2}$	$\sum_i N_i P_{c1i} + \sum_j M_j P_{c2j} + \sum_k O_k P_{c3k}$
Standard enthalpy of formation, H_f	$H_f - H_{f0}$	$\sum_i N_i H_{f1i} + \sum_j M_j H_{f2j} + \sum_k O_k H_{f3k}$
Standard enthalpy of vaporization, H_v	$H_v - H_{v0}$	$\sum_i N_i H_{v1i} + \sum_j M_j H_{v2j}$
Standard enthalpy of fusion, H_{fus}	$H_{fus} - H_{fus0}$	$\sum_i N_i H_{fus1i} + \sum_j M_j H_{fus2j} + \sum_k O_k H_{fus3k}$
Acentric factor, w	$\exp(w/a)^b - C$	$\sum_i N_i w_{i1} + \sum_j M_j w_{j2}$
Liquid molar volume, V_m	$V_m - d$	$\sum_i N_i V_{m1i} + \sum_j M_j V_{mj2}$
Surface Tension, σ	σ	$\sum_i N_i \sigma_{i1} + \sum_j M_j \sigma_{j2}$

Table 2.3: Adjustable Parameters in Group Contribution Models

<i>Adjustable Parameter</i>	<i>Value</i>
T_{m0}	147.45 K
T_{b0}	222.543 K
T_c	231.239 K
P_{c1}	5.9827 bar
P_{c2}	0.108998 bar ^{-0.5}
V_{c0}	7.95 cm ³ /mol
G_{f0}	-34.967 kJ/mol
H_{f0}	5.549 kJ/mol
H_{v0}	11.733 kJ/mol
H_{fus0}	-2.806 kJ/mol
a	0.4085
b	0.505
c	1.1507
d	0.01211

The application of group contribution based CAMD techniques rely on the availability of molecular groups and the estimated property contributions corresponding to each group. The properties that can be predicted based on the molecular structure alone are called primary properties and the properties that can be estimated as a function of primary properties and molecular structural data are called secondary properties (Constantinou & Gani, 1994).

2.6. Topological Indices and Property Prediction

The chemical structure of a molecule can provide a lot of information about the properties that it possesses. The information available from a structural formula includes (1) total number of atoms (2) number of each type of atoms and (3) the bonding between the atoms. These sets of information enable the structure to be represented in a graphical form (Kier & Hall, 1986). The representation of a molecule in the form of a graph is the first step in the development of topological indices. This would allow the conversion of the structural formula into indices and a potential opportunity to relate the structure to properties (Kier & Hall, 1986). Once the chemical structure is represented in the form of a graph, a number of graph theoretical matrices can be formed from the chemical structure. The most commonly used matrices in the formation of topological indices are adjacency matrix and distance matrix (Trinajstic, 1992). The vertex adjacency matrix can be represented as shown in eq. (2.12):

$$(A)_{ij} = \begin{cases} 1 & \text{if vertices } v_i \text{ and } v_j \text{ are adjacent} \\ k & \text{if vertices } v_i \text{ and } v_j \text{ are adjacent and edge } (v_i \text{ and } v_j) \\ & \text{is } k \text{- weighted} \\ 0 & \text{otherwise} \end{cases} \quad (2.12)$$

Here, $A(G)$ is the vertex adjacency matrix of the connected molecular graph G . This matrix is an $N \times N$ symmetrical matrix, where N is the number of vertices.

The topological indices can be calculated from the graph theoretical matrices by performing different operations over the matrix. The topological index of a molecular graph is a single number that can be used to characterize that graph. Therefore, this

number must have the same value regardless of the way in which the graph is labeled (Trinajstić, 1992). A topological index thus is a convenient way for representing the chemical constitution in the form of a number. The challenge in developing topological indices is that the descriptor should be able to form Quantitative Structure Property Relationships (QSPR) or Quantitative Structure Activity Relationships (QSAR). Randić and Basak (2001) suggested that, in order for a topological index (TI) to be of practical significance, it should possess a set of desirable attributes. The important qualities that make a meaningful TI include direct structural interpretation, correlation with at least one property, linearly independent, non-triviality, a basis on structural concepts, etc.

The topological indices are defined based on the topology of a molecule. These are developed based on the principles in chemical graph theory. Graph theory is a branch of mathematics that deals with objects that are connected (Wilson, 1986). The objects in the graph are called vertices, the lines used to connect the objects are called edges, and the diagram thus obtained is called a graph. The relationships developed pertaining to graphs have been extended to different disciplines. The principles in graph theory applied to analyze the consequences of connectivity in a chemical graph are termed as chemical graph theory (Trinajstić, 1992). Here, the sites may be atoms, molecules, molecular groups, etc. and the connection between those sites may be bonds, interactions etc. The branch of chemical graphs that represent the constitution of molecules are called molecular graphs. More details on applications of molecular graphs in molecular synthesis are given in section 2.8.

The molecular graphs are generally represented as hydrogen-suppressed graphs where only the molecular skeletons without hydrogen atoms (except for heteroatoms) are

used. Double and triple bonds are also not shown in the hydrogen-suppressed graph. The presence of hydrogen atoms and multiple bonds are handled in the general formulation of molecular indices. The difference between normal molecular structures and hydrogen-suppressed graphs is shown in figure 2.7. The first figure is the molecular structure of acetone and its hydrogen suppressed graph representation is shown in the second figure.

In the hydrogen-suppressed graph, the numbers 1, 2 and 4 represent the carbon atoms without hydrogen atoms on it and 3 represents the oxygen atom. The edges *a*, *b* and *c* represent the bonds connecting these atoms. Even though the double bond and single bonds are represented identically, the definition of bond indices is defined in such a way to take care of that difference.



Figure 2.7: Example of Hydrogen Suppressed Graphs

The objective in developing quantitative structure activity relationships (QSAR) and quantitative structure property relationships (QSPR) is to develop practical tools to relate the properties to chemical structure. The property of a molecule can only be explained in terms of the three dimensional aspects known as molecular topography such as shape, volume and surface area. However, the topographical characteristics are indeed related to the nature of individual atoms and the bonds between them. The effect of the bonding pattern is such that it controls the topography and thereby the properties.

Therefore, the properties also must be related to the topology of the molecule (Kier & Hall, 1986).

An exhaustive study has been conducted by Katritzky and Gordeeva (1993) with a variety of classical topological indices and geometric/electric descriptors for their ability to provide meaningful QSAR/QSPR relationships. It has been confirmed that, the classical topological indices give the best correlations for the determination of physical properties whereas a combination of topological indices and geometrical descriptors give the best quality regression expressions even though the relationships with topological indices alone also did not perform too bad. However, for most of the topological indices, no unambiguous criterion has been followed for their selection and verification (Gutman & Polansky, 1986). Therefore, many of them may contain the same kind of structural information with the difference being only in the scaling factor (Trinajstic, 1992).

Most topological indices are developed either from the adjacency matrix or from the distance matrix of the molecular graph (Trinajstic, 1992). However, since the distance matrix can be developed from an adjacency matrix (Bondy & Murty, 2008), the topological indices can be developed from the adjacency matrix alone. Some of the common topological indices are described in the following sections.

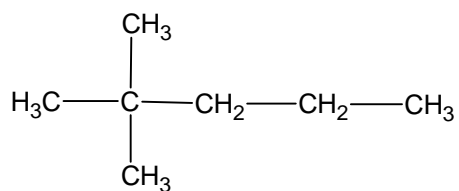
2.6.1. Connectivity Indices

The basic assumption employed in the connectivity index method is that, the structural formula of a molecule has enough information for relating it to its properties. Therefore, the efforts to obtain non-empirical expressions for such index values are logical. In one of the initial works, Randic (1975) found that in the alkane skeletons, the

number of adjacent carbon atoms to one specific atom can give a description of the branching of the molecule. In that work, he proposed to use a molecular descriptor called the delta value. The delta value is the count of formally bonded carbon atoms. The delta value is obtained from the individual atomic valencies of the atoms that form the bond. The product of the atomic valencies is raised to the power of -0.5 to obtain the delta value.

The sum of all delta values in the molecule provide an index associated with that molecule. The larger the branching in a structure, the lower will be the branching index corresponding to that structure because of the inverse relationship.

The following example illustrates the calculation of the branching index value for a 3,3 dimethyl pentane molecule:



The structure is described with a molecular skeleton with the count of all bonded carbon atoms in figure 2.8:

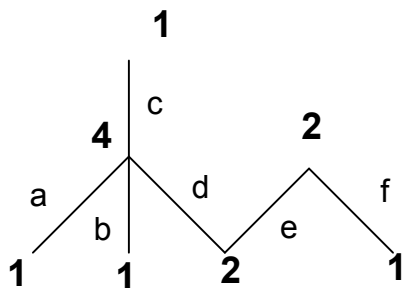


Figure 2.8: Molecular Skeleton of 3,3 Dimethyl Pentane

Here, a, b, c, d, e, f represents different bonds. According to the definition above, the delta value corresponding to each bond can be calculated as follows:

$$\text{Bond } a : (1 \cdot 4)^{-0.5} = 0.5$$

$$\text{Bond } b : (1 \cdot 4)^{-0.5} = 0.5$$

$$\text{Bond } c : (1 \cdot 4)^{-0.5} = 0.5$$

$$\text{Bond } d : (2 \cdot 4)^{-0.5} = 0.3535$$

$$\text{Bond } e : (2 \cdot 2)^{-0.5} = 0.5$$

$$\text{Bond } f : (2 \cdot 1)^{-0.5} = 0.707$$

$$\begin{aligned} \text{Branching Index of 3,3 dimethyl pentane molecule} &= 0.5 + 0.5 + 0.5 + 0.3535 + 0.5 + 0.7 \\ &= 3.0605 \end{aligned}$$

The branching number was found to correlate with properties like boiling point, Kovats constants and a calculated surface area (Kier & Hall, 1986).

However, additional descriptors developed based on delta values have the limitation that they would not differentiate among saturated and unsaturated bonds.

Therefore, a new concept was developed by Kier and Hall (1976). The new structural descriptor is termed as the valence delta (δ_v) which is based on the explicit counting of each bond to a nearby atom and it is estimated as follows:

$$\delta_v = Z^v - h \quad (2.13)$$

Here, Z^v of an atom is the count of all adjacent bonded atoms and all *pi* and lone pair electrons and h is the number of hydrogen atoms bonded to that atom.

While dealing with high atomic weight atoms, the effect of non-valence core electrons to the atomic size and properties must be considered because such core electrons also have a significant role in influencing the properties. For that, the valence delta is redefined for an atom of atomic weight Z as follows:

$$\delta^v = \frac{(Z^v - h)}{(Z - Z^v - 1)} \quad (2.14)$$

The valence delta values for heteroatoms at higher oxidation states will be different from the values given by the above formula. Some empirical values for a few atoms at different oxidation levels have been estimated and are available in literature (Kier & Hall, 1986).

If i and j are the atoms involved in the bond, then, bond indices, β^k are defined through the pair of valence delta values (Kier & Hall, 1986).

$$\beta^k = \delta_i^v \delta_j^v \quad (2.15)$$

The zero'th order connectivity index of an atom has been estimated from the individual valence deltas and the first order CI is formed from the possible bonds present in the molecule as follows (Kier & Hall, 1986):

$${}^v\chi^0 = \frac{1}{\sqrt{\delta_i^v}} \quad (2.16)$$

$${}^v\chi^1 = \frac{1}{\sqrt{\beta^k}} \quad (2.17)$$

Zero order and first order connectivity indices are the most commonly used molecular descriptors. However, higher orders of connectivity indices can be calculated by following the same methodology. The n^{th} order connectivity index is calculated using eq. (2.18):

$${}^v\chi^n = \frac{1}{\sqrt{(\delta_v^i \dots \delta_v^n)}} \quad (2.18)$$

The zero order connectivity indices (CI) for a molecule are obtained by summation of the CI values of each atom:

$$\left({}^v\chi^0\right)_{molecule} = \sum_1^n \left({}^v\chi^0\right)_n \quad (2.19)$$

The first order CI for a molecule is obtained as the sum of CI's of all edges:

$$\left({}^v\chi^1\right)_{molecule} = \sum_1^n \left({}^v\chi^1\right)_n \quad (2.20)$$

It is to be noted that, if the 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 by the bonds between two separate groups is not represented in the expression. To account for that, an additional term for the bond between different groups has to be included in the expression (Gani *et al.*, 2005).

$$\left({}^v\chi^1\right)_{group} = \sum_1^k \left(\frac{1}{\sqrt{\beta_{\text{internal bonds}}^k}} \right) + \sum_m \left(\frac{0.5}{\sqrt{\beta_{\text{bonds out of groups}}^m}} \right) \quad (2.21)$$

Here, k is the number of bonds inside the group for which the expression is written and m is the number of free bonds the group has.

2.6.2. Edge Adjacency Index

The edge adjacency index (ε) is a topological index developed by Estrada (1995a). The reported correlation coefficient ($R > 0.99$) of ε in QSPRs to estimate molar volume is significantly higher than any of the available group contribution methods. The calculation of ε index is described below.

The development of the vertex adjacency matrix has been explained in section 2.6. Similar to that, an edge adjacency matrix can also be developed. Two edges are

adjacent if one vertex in a chemical graph is incident (having one common vertex) to both the edges. If there are m edges in a graph and g_{ij} are the elements in that graph, the edge adjacency matrix, $E = [g_{ij}]_{m \times m}$ can be defined as shown in eq. (2.22):

$$g_{ij} = \begin{cases} 1 & \text{If } e_i \text{ and } e_j \text{ are adjacent} \\ 0 & \text{Otherwise} \end{cases} \quad (2.22)$$

The edge degree $\delta(e_k)$ has been defined in eq. (2.23):

$$\delta(e_k) = \sum_i g_{ik} \quad (2.23)$$

This can be calculated as the sum of elements of k^{th} row in the matrix E. Now, edge adjacency index can be calculated using eq. (2.24):

$$\varepsilon = \sum_l [\delta(e_i) \delta(e_j) \cdot k]^{-1/2} \quad (2.24)$$

where the sum is over all l adjacent edges in the graph. Here, k is a constant, which is defined in eq. (2.25):

$$k = \begin{cases} 1 & \text{if } e_i \text{ and } e_j \text{ are adjacent} \\ 0 & \text{Otherwise} \end{cases} \quad (2.25)$$

If the graph contains heteroatoms, it is necessary to account for the differences in bonds formed between heteroatoms and carbon from the other types of bonds. In that case, the values of g_{ik} are replaced with K_{C-X} , which are the values corresponding to the bonds between carbon and the heteroatom (Estrada, 1995b). The K_{C-X} parameters are related to the resonance integral associated with the bond between the heteroatoms and the carbon atom (R. Daudel *et al.*, 1959). Different values have been reported in the literature for K_{C-X} . The values reported by Ortiz and Perez (1982) are shown in table 2.4:

Table 2.4: Values of K_{C-X} Parameters

<i>C-X Bond</i>	K_{C-X}
C-N	0.9
C-O	0.8
C=O	1.6
C-S	0.7
C-F	0.7
C-Cl	0.4
C-Br	0.3

2.6.3. Shape Indices

The shape index is a molecular descriptor used for the quantification of the molecular shape. Depending on different aspects of the shape, shape indices of different orders have been developed. Shape index of order one is defined as shown in eq. (2.26):

$${}^1K = \frac{n(n-1)^2}{({}^1P)^2} \quad (2.26)$$

1P is the number of paths of length 1. The other shape indices can also be calculated in the similar manner.

2.6.4. Wiener Indices

The Wiener number is introduced as the path number, which is the number of bonds between all pairs of atoms in an acyclic molecule. The main significance of this index is that, this was the first time the significance of paths in a molecular skeleton had been recognized (Wiener, 1947). The Wiener number, W is defined as one half of the off-diagonal elements of the molecular distance matrix:

$$W = 0.5 \sum_{k=1}^N \sum_{l=1}^N D_{kl} \quad (2.27)$$

where D_{kl} is the off-diagonal elements of the distance matrix, D .

Together with other molecular descriptors, the Wiener number can make predictions on alkane properties like boiling points, heats of formation, heats of vaporization, molar refractions and molar volume.

2.6.5. The Hosoya Topological Index

The Hosoya topological index is defined by the count of all possible patterns of considering k disjoint bonds in a molecule (Hosoya, 1971):

$$Z = 1 + Z_1 + Z_2 + \dots + Z_K \quad (2.28)$$

Here, Z_1 is the number of bonds in a graph, Z_2 is the number of pairs of disjoint bonds, Z_3 is the number of triples of disjoint bonds and so on. The disjoint bonds are any two or more bonds in the structure for which there is no incident vertex.

A variety of other connectivity indices is available in many published works. A good review and the classification and applications of different topological indices can be found in Trinajstić (1992) and Balaban (2001).

2.7. Connectivity Indices and GC⁺ Method

The GCM can provide fairly accurate estimates of the properties of the molecules if the group contribution values of all the building blocks are known. However, if a group, whose property contribution is not available, makes up at least one part of the molecule the property estimation cannot be completed. In practice, this is a very common frustration encountered while using GCM for property estimation or in reverse problems because, property contributions of many common molecular groups are not available in literature. So, recent works on the correlation between connectivity indices and some physical properties can be used in such situations (Gani *et al.*, 2005).

In a recent work to correlate the CI's to physical properties, the following pure component property model was proposed by Gani *et al* (2005):

$$f(Y) = \sum_i (a_i A_i) + b({}^v \chi^0) + 2c({}^v \chi^1) + d \quad (2.29)$$

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. The property functions are defined the same way as in the GCM. The values of the constants are given in the appendix (Table A.6).

It is to be noted that the constants in the above equation are regressed using the same pure-component property data used by Marrero and Gani (2001) in their group contribution model development. Here, as the derived equation is on the atomic level, the expected accuracy of prediction is less than the group contribution model. So, the CI model is used only for deriving the property contributions of groups not available in existing group contribution methods.

Since the property functions are defined the same way in both GCM and CI methods, the formulation of a combined approach is straightforward. The combined GC-CI model, known as the GC⁺ model, has been written as follows (Gani *et al.*, 2005):

$$f(Y_m) = \sum_i (a_{m,i} A_{m,i}) + b({}^v\chi^0)_m + 2c({}^v\chi^1)_m \quad (2.30)$$

$$f(Y^*) = \left(\sum_m n_m f(Y_m) \right) + d \quad (2.31)$$

$$f(Y) = \left(\sum_i N_i C_i \right) + f(Y^*) + \left(\sum_s N_s C_s \right) + \left(\sum_t N_t C_t \right) \quad (2.32)$$

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.

2.8. Molecular Signature Descriptors

2.8.1. Current Status in Inverse Design

The inverse design of identifying the molecules from property constraints is a relatively new problem from a chemical engineering perspective. In this section, the different methodologies developed for designing molecules with a set of target properties will be reviewed. It should be noted that, the current methodologies are very specific to certain classes of property models and even in such restricted situations, the accuracy and computational expenses requires a lot of improvement. Most of the existing methods used group contribution based approaches for solving the inverse design problem (Achenie *et al.*, 2003). A number of recent publications have used the group contribution based techniques to solve for different classes of molecular design problems (Sahinidis *et al.*, 2003; Achenie & Sinha, 2004; Eljack & Eden, 2008; Chemmangattuvalappil *et al.*, 2009). However, the suitability of group contribution methods for molecular design is limited because of the following reasons:

1. It is not always possible to find a suitable correlation between the molecular groups and properties
2. Not all possible atomic arrangements are represented in GCM
3. Many group contribution models have limited ranges of accuracy

As mentioned before, the representation of a molecule in the form of a graph can provide lot of information about its properties through the use of molecular descriptors.

Molecular descriptors are operators developed from the molecular graph to characterize the properties of the molecule. The numerical value obtained after performing the operations suggested by the descriptor on the molecular graph can generally be used to correlate and predict physical properties and biological activities (Faulon *et al.*, 2003b).

There are thousands of molecular descriptors available and that makes it difficult to select the appropriate one(s) for a specific problem. A lot of work has been done in molecular design using topological indices as structural descriptors (Baskin *et al.*, 1990; Gordeeva *et al.*, 1990; Kvasnicka & Pospichal, 1990; Kier *et al.*, 1993a; Skvortsova *et al.*, 1993). In all these approaches, the descriptors' structural features have been used to generate the feasible molecular structures. However, the inverse relationships between the topological indices generally do not provide a unique molecular graph (Trinajstic, 1992). Therefore, the degeneracy in these approaches is very large. In addition, most topological indices exhibit highly non-linear functional dependence on the elements of the vertex-adjacency matrix (Raman & Maranas, 1998). Because of that, obtaining a global solution when employing mathematical programming techniques is difficult.

The techniques developed recently for obtaining unique molecular structures from the molecular descriptors use stochastic approaches. In the algorithms developed by Venkatasubramanian *et al.* (1994, 1995), and Sheridan and Kearsley (1995), a genetic algorithm based approach has been introduced for large scale molecular design. Genetic algorithms are stochastic optimization methods based on the Darwinian model of evolution. In summary, genetic algorithms identify the population of best candidates from an earlier population following the rules of crossover and mutation and thus producing the best offspring for the next generation (survival of the fittest). A fitness function based

on the target properties has been used to evaluate the fitness of the candidate solutions. This approach is expected to identify better offspring after each generation and eventually end up with the optimal solution. These approaches were the first efficient methodology for solving large-scale molecular design problems. They have the ability to locate optimal designs for those problems with multiple target specifications. However, because of the heuristic nature of the developed algorithms, there is no guarantee that a solution will be obtained after running the algorithm. In addition, even though these algorithms can obtain a near optimal solution very fast, the efficiency is very limited in obtaining the final solution.

Algorithms based on Monte Carlo simulations (Faulon, 1996; Kvasnicka & Pospichal, 1996) have also been published. A simulated annealing based algorithm has been published by the group of Kokossis (Marcoulaki & Kokossis, 1998) for the design of refrigerants and liquid-liquid extraction solvents. One advantage of simulated annealing over other stochastic optimization methods is that, it can provide probabilistic guarantee on the quality of the final solutions. These algorithms can theoretically provide the solution in polynomial time. A variety of other papers have been published later using stochastic techniques, however, the reconstruction of molecular structures using deterministic techniques has rarely been attempted.

The notable contributions that use deterministic techniques for the exhaustive generation of molecular graphs corresponding to the predefined molecular descriptors have come from Kier's group (Hall *et al.*, 1993a; Hall *et al.*, 1993b; Kier *et al.*, 1993b) and from Skvortsova's group (Skvortsova *et al.*, 1993). The works from the former group compute the possible degree sequences that match the paths of the target descriptors up to

the length of two that can track degree sequences up to length three. These contributions applied the developed techniques to chi-indices. Once the different paths of length two are obtained, the degree sequences corresponding to all the molecular structures will be generated using an isomer generator. Only those structures that match the path of length 3 (which is the maximum path length that can be tracked using this technique) are accepted as the final solutions. In the contribution from Skvortsova's group, apart from the degree sequence, an edge sequence also has been generated, which simultaneously formed the input to the isomer generator to build the exhaustive list of corresponding structures. The edge sequence can estimate the number of edges between each combination of atoms. This can decrease the degeneracy of the solutions significantly. However, the descriptive features in these methods do not always produce feasible or unique structures.

In more recent works (Raman & Maranas, 1998; Camarda & Maranas, 1999), methodologies have been developed to incorporate topological indices within an optimization framework. In the work by Raman and Maranas (1998), many hydrocarbon properties are correlated with connectivity indices and shape indices of different orders. In this work, the molecular structure has been represented in the form of a vertex adjacency matrix that can completely explain the molecular interconnectivity. Even though the actual topological indices used in this work are non-linear, the matrix representation has been used to systematically transform them into linear form. A mixed integer linear program (MILP) formulation has been formed which ensures that a global optimum solution can be achieved. However, its application has been limited to the design of alkyl structures. In the work by Camarda and Maranas (1999), nonlinearities

due to the expressions for connectivity indices led to MINLP formulations, which make the solution methodology computationally expensive and susceptible to local optima traps.

Another important inverse problem solution technique had been developed by forming a target scaffold (Garg & Achenie, 2001) for drug design. This is the first attempt to apply mathematical programming techniques in the area of drug design. In this work, a target scaffold based on a drug molecule has been used to generate a QSAR and the inverse problem has been solved for the best values of selectivity by changing the substituents on the scaffold. The limitations of this approach was that it was not able to provide nonintuitive solutions because the scaffold limits the type of molecules obtained as solutions. However, this approach was effective in controlling the combinatorial explosion.

In a later contribution, a fitness function was directly incorporated into an optimization framework (Siddhaye *et al.*, 2004). However, in many formulations, a globally optimal solution cannot be guaranteed in this approach. In another work, second order connectivity indices had been used for the design of value added soybean oil products (Camarda & Sunderesan, 2005). The interesting aspect of this work is that, the highly non linear second order connectivity indices have been represented with the exact linear equivalent expressions using Glover transformation. Glover transformation is a technique used in non-linear integer programming to represent non-linear expressions in terms of linear equations that captures the essential non-linearities of the original problem. By applying Glover transformation, the non-convex terms have been converted into products of binary variables. Even though this approach eliminates the possibility of

local minima traps by avoiding non-convex terms, the computational requirements are relatively high. To decrease the computational complexity, an approach known as ‘templating’ has been applied in this work. In templating, a portion of the vertex adjacency matrix has been predefined to control the generation of a large number of structures. However, because of templating, the generation of non-intuitive solutions may be eliminated.

In some recent works, heuristic methods have been used in order to handle the non-linear constraints. (Lin *et al.*, 2005; Eslick *et al.*, 2009; McLeese *et al.*, 2010). These methods cannot ensure a global optimum solution. However, in order to ensure that the solution is *near optimum*, Tabu search methods have been employed while solving the problem. In Tabu search, a library will be generated that keeps track of the recently generated local optima solutions that will prevent the generation of the same local minima solution as the search proceeds (Lin *et al.*, 2005). Even though this technique ensures better quality of the solutions, the optimum solution can never be guaranteed because of the non-linear constraints.

Apart from these limitations, the above-mentioned techniques can be used only when the QSAR/QSPRs are based on one topological index. For many properties, the QSAR/QSPRs are formed based on more than one topological index. In addition, the topological indices required to form the QSAR/QSPRs may be different for different properties of interest. For instance, the QSPR for one property target may be based on connectivity indices and the second property target may be based on shape indices. Since different topological indices are formulated using different mathematical expressions, there is no standard way to combine everything on a common platform and solve it all

simultaneously. Therefore, the current techniques that employ QSAR/QSPR models in reverse problem formulations can handle those property models with one topological index. Different topological indices have to be formulated using different mathematical transformations. In addition, the degeneracy of the solutions in all these methodologies is very high. That means, for a specific solution, there could be many possible molecular structures. The degeneracy increases with the size of the molecules in the solution. Here, we are looking for a computationally efficient algorithm that can simultaneously incorporate different topological indices based on QSAR/QSPRs. The recent works on molecular signature descriptors (Visco *et al.*, 2002; Faulon *et al.*, 2003a; Faulon *et al.*, 2003b; Weis *et al.*, 2005) provides a convenient way to represent a variety of TIs as linear combinations of molecular signatures. Therefore, this approach has the potential to develop an efficient methodology for the design of molecules with QSARs/QSPRs related to diverse TIs.

2.8.2. Development of Molecular Signature

The concept of molecular signature (Visco *et al.*, 2002; Faulon *et al.*, 2003b) is significant for the reverse problem formulation framework because, it forms a finite set of not highly correlated descriptors based on the molecular structure from which all other TI's can be calculated.

The molecular signature is a systematic way of representing the atoms in a molecule using the extended valancies to a pre-defined height. The systematic procedure for the identification of the signature of a molecule developed by Visco *et al.*, (2002) is explained in the next section.

One of the characteristics that make the molecular signature descriptors unique among other molecular descriptors is that the building blocks of the molecular signatures complement each other. 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 subgraph of G containing all atoms that are at a distance h from x . This canonical representation can be obtained by the following systematic procedure:

1. All atoms (vertices) in the graph are labeled in a canonical order starting with atom x
2. For the atom x for which the atomic signature is to be constructed, all atoms and bonds will be shown up to the height h in the sub graph ${}^hG(x)$.
3. Construct the tree that spans over all the edges in the sub graph. The root of the tree is the atom x itself. The tree is constructed one layer at a time up to level h . The first layer of the tree are the nearest neighbors of atom x , the second layer of the tree consists of all the neighbors of the vertices in layer 1 except the atom x . In general, when the tree has been constructed up to level $h-n$, then, layer $h-n+1$ will be constructed considering each vertex of layer $h-n$. All vertices in the tree are labeled and colored with the necessary coloring function. The vertex color of a graph is the assignment of a unique description provided to its vertices in order to distinguish among different groups of atoms. The coloring function will be selected based on the type of the molecule. Typical coloring functions include the type of atoms, type of bonds and valency. It is possible to have one vertex more than once in the graph. However, no edge should be repeated in the same graph.

4. The signature can be written by reading the tree from the atom x . The child level vertices will be enclosed in parenthesis at each level. The vertex color must be written along with the vertices in each level. After each level, the next level must be followed until the signature reaches the required height. While writing the signatures, all the neighbors including the root atom must be written.

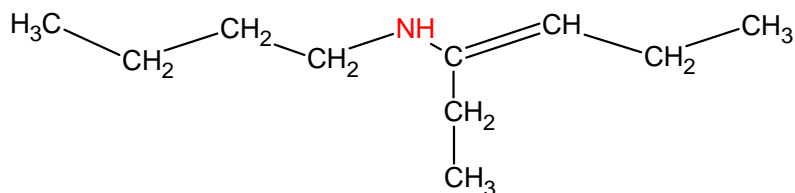
One example of the construction of atomic signature is shown in figure 2.9. Here, the stepwise procedure for obtaining the atomic signature of atom N (nitrogen) up to height 3 in a molecule is illustrated.

In the first step, the atoms are labeled to distinguish between them when writing the molecular signatures in the later stages.

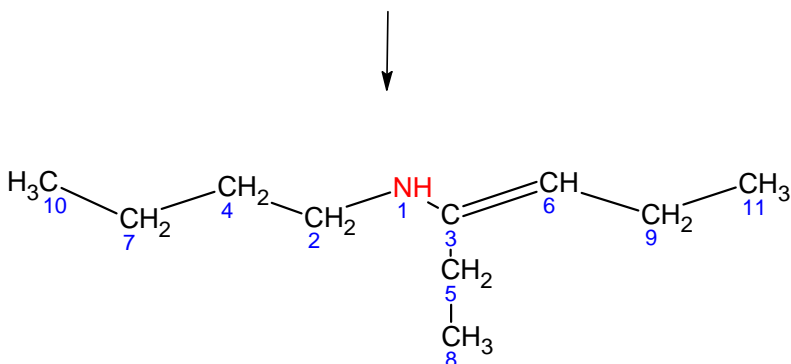
In the second step, all the atoms at height 3 from atom N are extracted. In other words, the neighbors of height three includes the atoms bonded to N (say y), the atoms bonded to all atoms in y (say z) and all the atoms bonded to all atoms in z .

In the third step, the molecular groups are replaced with vertices. All the vertices are colored with the atom type, valency and the type of bond. Here the atoms types are C and N. The different carbon types are distinguished with their valancies and the type of bond. Note that the bond type has been retained in the canonical representation in step 2. In the final step, the signatures of different heights have been formed by reading the tree starting from the root N atom. The atomic signature of height zero is the root N atom itself. The atomic signature of height one is the root atom followed by its nearest neighbors (in this case, two carbon atoms) enclosed in parenthesis. In signature of height two, the neighbors of the carbon atoms (including the root N atom) are listed in the

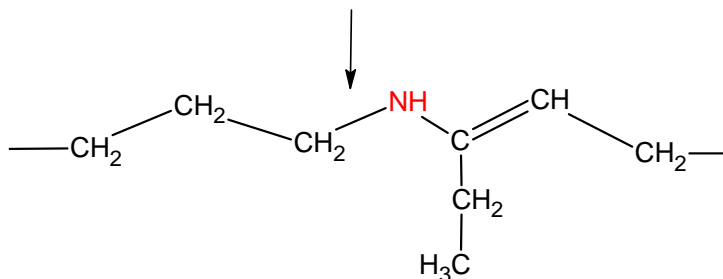
parenthesis. Finally, signature of height three is obtained by adding the neighbors of the atoms in the previous layer. While writing the signatures, the vertices at the different levels have been color coded for clarification of the levels.



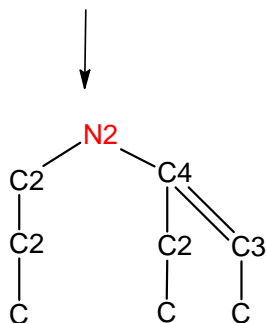
Molecular structure



Step 1



Step 2



Step 3

$${}^0\sigma(x) = N$$

Height 0

$${}^1\sigma(x) = N2 (CC)$$

Height 1

Step 4

$${}^2\sigma(x) = \text{N2 (C2 (NC) C4 (=CCN))} \quad \text{Height 2}$$

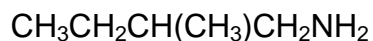
$${}^3\sigma(x) = \text{N2 (C2 (N2(CC) C1(C)) C4 (=C3(=CC) C1(C) N2(CC)))} \quad \text{Height 3}$$

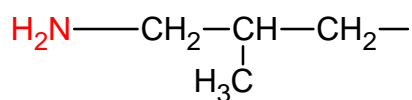
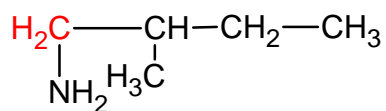
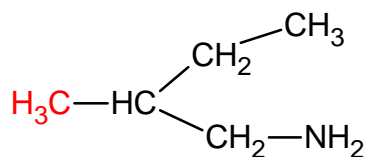
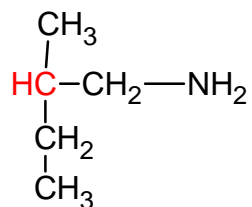
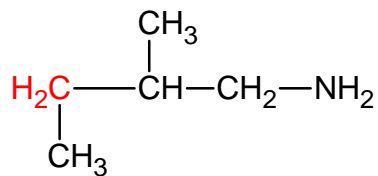
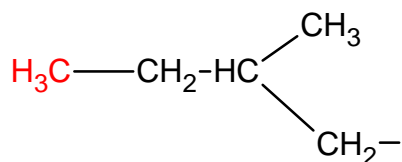
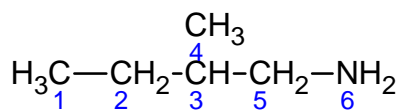
Figure 2.9: Atomic Signatures up to Height 3

It is clear that the set of atomic signatures up to a given height is of finite size. So, any molecule can be represented by its coordinates in a vectorial space where the base vectors are its atomic signatures. Thus, the signature of a molecule is defined as the linear combination of atomic signatures (Visco *et al.*, 2002; Faulon *et al.*, 2003b). 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:

$${}^h\sigma(G) = \sum_{x \in V_G} {}^h\sigma_G(x) = \sum_{i=1}^{{}^hK_G} {}^h\alpha_i {}^h\sigma_G({}^hX_i) \quad (2.33)$$

A simple example is presented to show molecular signatures of different heights using eq. (2.33). In the first step, the subgraphs have been generated for all the atoms that produce signatures up to the required height. Then, individual atomic signatures for all the atoms have been generated. Finally, eq. (2.33) is applied to generate the molecular signature.





$${}^0\sigma = 5C + N \quad \text{Height 0}$$

$${}^1\sigma = 2C1(C) + C2(CC) + C3(CCC) + C2(NC) + N1(C) \quad \text{Height 1}$$

$${}^2\sigma = C1(C2(CC)) + C2(C1(C)C3(CCC)) + C3(C1(C)C2(CC)C2(NC)) + \\ C1(C3(CCC)) + C2(C3(CCC)N1(C)) + N1(C2(NC)) \quad \text{Height 2}$$

$${}^3\sigma = C1(C2(C1(C)C3(CCC))) +$$

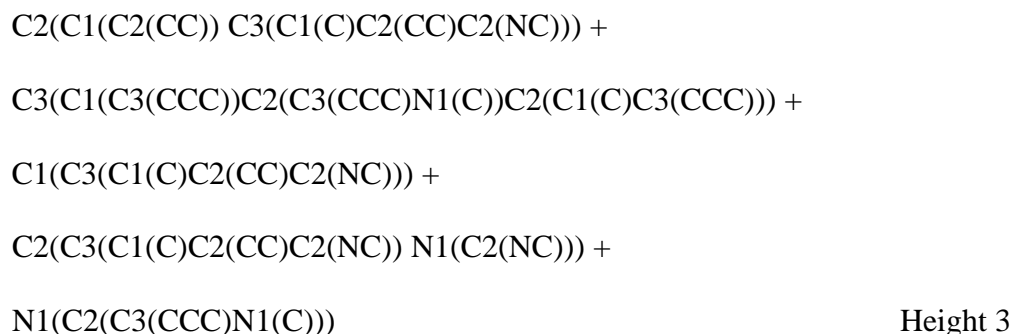


Figure 2.10: Molecular Signature Tree

2.8.3. Application of Signature Descriptors in Property Prediction

The atomic signature concept is useful in QSAR/QSPR studies because of its applicability in defining many topological indices. Consider a molecule G with known atomic signature up to height h as defined in section 2.6.1. Suppose the number of vertices in an intermediate layer k of graph G is ${}^kV(x)$. Note that $k < h$. It has been observed that many topological indices can be computed from the molecular signature of height h , where $h \leq n$, where n is the number of atoms computed from signature of height zero (Visco *et al.*, 2002; Faulon *et al.*, 2003b). More discussion of the calculation of different topological indices from signatures is given in chapter 4. Therefore, it can be concluded that, the QSAR/QSPR relationships can be re-written in terms of signatures of different heights no greater than the number of atoms involved. Once the forward topological index-property relation has been identified, the next step is to develop an algorithm for obtaining the constituent atomic signatures corresponding to a specific property, which will be explained in chapter 4.

2.9. Flowsheet Property Model

A flowsheet property model can quantify the efficiency of different processing routes from raw materials to products. In a recently published work, a flowsheet property model has been proposed to estimate the energy consumption of a unit operation (d'Anterrosches & Gani, 2005). In this work, a flowsheet is considered as a combination of different process groups. Similar to the molecular groups used in group contribution models, these groups also have one or more free bonds, which can be used to link with other process groups. In this way, a variety of unit operations can be represented using the collection of process groups and a variety of flowsheet properties can be calculated, albeit theoretically, based on the contributions of each process group to the flowsheet property.

In order to apply a group contribution type approach for the design of a flowsheet, certain rules have to be followed. Most of the connectivity principles are synonymous with those used in the design of molecules from molecular groups. However, the property for which the models are developed must provide a measure of the performance of the operations in a flowsheet. In addition, those properties should be expressed as a function of the contributions of each unit operation of the process. The generic model for a group contribution based property estimation model has been developed by d'Anterrosches and Gani (2005):

$$f(P) = \sum_{k=1}^{NG} pos_k \cdot a_k \quad (2.34)$$

where $f(P)$ is a function of the property that can be described using the group contribution model, a_k is the regressed contribution of various process groups, NG is the number of process groups and pos_k is the topology factor.

Even though this method can theoretically be applied for a variety of unit operations and for a variety of flowsheet properties, the models are currently only available for distillation systems and for the estimation of energy index. In the available model for the calculation of a energy index (d'Anterrosches & Gani, 2005), the energy consumption of a distillation column that separates a single inlet stream into two product streams can be estimated. This model has been developed based on the driving force based model for distillation (Bek-Pedersen & Gani, 2004). The driving force, D_{ij} is defined as:

$$D_{ij}(x_i, y_i, \alpha_{ij}) = \frac{x_i \alpha_{ij}}{1 + x_i (\alpha_{ij} - 1)} - x_i = y_i - x_i \quad (2.35)$$

where, x_i and y_i are the mole fractions of the component i and α_{ij} is the relative separability of the component i with respect to component j .

The conclusions obtained from the driving force based approach for the design of distillation systems are (Bek-Pedersen & Gani, 2004):

1. The driving force is inversely proportional to the energy consumption
2. In a system where distillation is employed to separate a number of components, the separation with the maximum driving force must be performed first.

Bek-Pedersen and Gani (2004) have shown that, if the driving force of a distillation process can be fixed, the other design parameters such as feed plate location,

optimum reflux ratio, etc. can be obtained corresponding to the optimum process performance. If the driving force is the input variable, a property model to predict any of the flowsheet properties can be considered as a component independent model. Theoretically, any other unit operation can also be considered based on the driving force. If the driving force corresponding to the unit operation can be identified and the parameters can be regressed for a variety of flowsheet properties, the design can be conducted based only on driving force independent of the identities of the components involved.

The available property model for the estimation of energy index for a distillation system has been given in eq. (2.36). This model can be applied when the process groups, the driving force that can be obtained once the component identities are known, and the group contributions are available:

$$E = \sum_{k=1}^{NG} \left(\frac{1+p_k}{d_{ij}^k} \times a_k + A \right) \quad (2.36)$$

where, NG is the total number of process groups, d_{ij}^k is the maximum driving force of process group k , a_k is the contribution of process group k , A is a constant, which is different for different unit operations and E is the energy index. p_k is a topology factor defined in eq. (2.37):

$$p_k = \sum_{i=1}^{nt} \overline{D}_i \quad (2.37)$$

where, nt is the number of separation tasks that should be performed before task k in the ideal case and \overline{D}_i is the maximum driving force of task i .

Similar to the way the group contribution parameters are developed, the property contributions of various process groups have been estimated by fitting experimental and simulation data to the regression expressions. Currently, distillation columns separating up to five component mixtures to products with different specifications are available (d'Anterrosches & Gani, 2005).

2.10. Summary

This chapter provided a brief discussion of the field of product design and the motivation and challenges put forward by integrated process and product design problems. The chemical engineering community is now concentrating on developing methodologies to identify products corresponding to optimum process performance. For the consumer, the properties of the output material rather than its chemical composition define the suitability of that material. Therefore, design based on properties is a smart approach to ensure customer satisfaction.

Since the desired algorithms should be based on the properties of not yet defined molecules, the models employed to measure properties and use them in designing molecules are important. The classification of different types of properties and property models has been discussed. The general mathematical formulations of different classes of product design problems have been explained and the challenges put forward by integrated process and product design problems were discussed. Statistical design of experiments is the traditional way to identify appropriate experiments for obtaining

optimum product formulations. However, the combinatorial explosion caused by the huge number of combinations of building blocks demands novel techniques to be employed prior to conducting actual experiments. To design products based on properties, group contribution methods (GCM) are proven techniques and a detailed description of GCM and the current developments in that field has been presented. The recently developed connectivity index based property estimation methods are described next to account for molecular groups not described by GCM. For the determination of properties that require detailed structural information, the topological index based QSAR and QSPR relationships are useful. An overview of a number of the most commonly used topological indices has been provided and the computational complexities encountered when applying them for solving inverse problems has been discussed. Even though the topological indices can be used to predict a number of pure component properties from the molecular structure, there are no efficient algorithms to apply them in reverse problems. In addition, to account for the computational complexity of the topological index based expressions to be applied in inverse problems, there must be a way to represent different topological indices on a common platform. The recently introduced concept of molecular signatures provides a unique molecular descriptor to describe different property-structure relationships on a common platform. Detailed descriptions of the generation of atomic signatures have been provided and the calculation of molecular signature from molecular structure was illustrated. Finally, a recently introduced algorithm to calculate flowsheet properties using a group contribution based approach has been discussed.

The different property prediction models presented in this chapter provide excellent tools to calculate the pure component properties of different classes of molecules. However, there are no efficient algorithms for incorporating these property models in reverse problems. In a typical integrated process and product design problem, it is required to generate the potential molecular structures corresponding to the property targets identified during the process design stage. Apart from the process design targets, a number of environmental, health and safety constraints are also important while designing a molecule for an industrial process. Currently, there are reliable property models that can predict such properties from the molecular structure. Therefore, the objective of this dissertation is to develop different algorithms corresponding to different types of molecular design problems by making use of the available property models without sacrificing their accuracy. Since the molecular signature descriptors can represent a number of topological indices on a common platform, it can provide a useful tool in integrated process and molecular design problems.

3. Basics of Computer Aided Molecular Design

3.1. Computer Aided Molecular Design Framework

The design of molecules corresponding to a set of desirable characteristics has traditionally been considered as an iterative approach. The process involves an exhaustive search among a large number of candidate molecules. A generalized framework of this approach is given in figure 3.1. (Venkatasubramanian *et al.*, 1994):

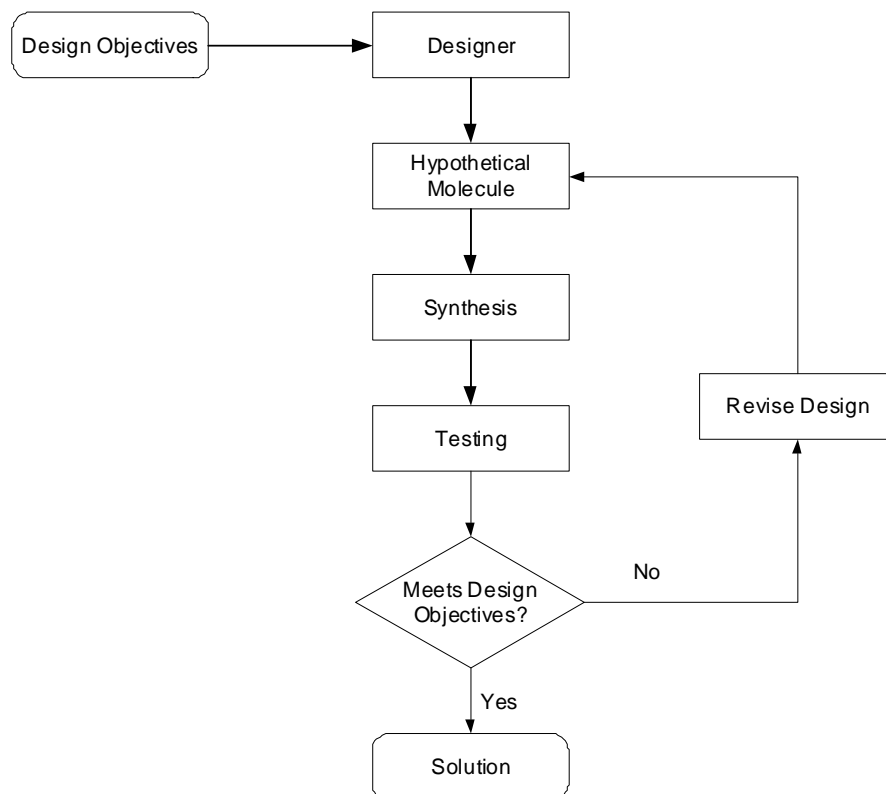


Figure 3.1: Iterative Molecular Design (adapted from Venkatasubramanian *et al.*, 1994)

Computer Aided Molecular Design (CAMD) techniques are efficient alternatives to the traditional iterative approach for molecular design. A computer aided molecular design problem can be considered as the reverse of a property prediction problem. The designer is required to enumerate the possible molecular structures based on the desired property targets (possibly identified by solving the process design part) and a set of molecular building blocks. This part can be considered as the pre-design stage (Harper *et al.*, 1999). In the second step, classified as the CAMD design step by Harper and Gani (2000), the feasible molecules that can satisfy the property targets are generated and tested against the specifications. A very good overview of of the many different techniques used for solving this kind of problem is available in a recently published book by Achenie *et al.* (2003). Since the design involves the application of a variety of classes of property models, the identification of suitable candidate/candidates has been considered to be a multilevel problem. The multilevel approach for product design suggested by Harper and Gani (2000) is shown in figure 3.2. Here, through successive steps of generation and screening against the design specifications a set of candidate molecules is identified. In the first step, the property targets, which could have been identified through process design, and a set of molecular building blocks form the input to the first stage. Here, CAMD tools based only on first order groups are used to identify the molecules that meet the property targets. The rules regarding the feasibility of molecular structure can be used to prevent combinatorial explosion. In the second stage, the CAMD techniques based on higher order group contribution methods are applied to eliminate infeasible candidates from the compounds identified in the first stage. In the third stage, the molecular structures need to be represented on an atomic scale. QSAR

and QSPR based property estimation techniques will be used in this stage. The short listed structures after the third stage will be analyzed using three-dimensional representations. Here, more rigorous analysis of the short listed structures will be carried out that includes database search, process simulation analysis, molecular modeling tools, etc. to differentiate between *cis/trans* and *R/S* isomers. The final step is termed as the post design step. The purpose of this stage is the verification and analysis of the factors, which are not predicted by CAMD tools. This analysis includes supplier database searches to verify the identified candidate molecules are commercially available at a reasonable price. In addition, database analysis provides valuable experimental and environmental data used to verify the results obtained through the multistage analysis. The environmental information and federal regulations are also important factors before making the final selection.

Generally, CAMD methods and tools work at the macroscopic level where the molecular structures are represented using groups (Gani, 2004b) or topological indices (Camarda & Maranas, 1999).

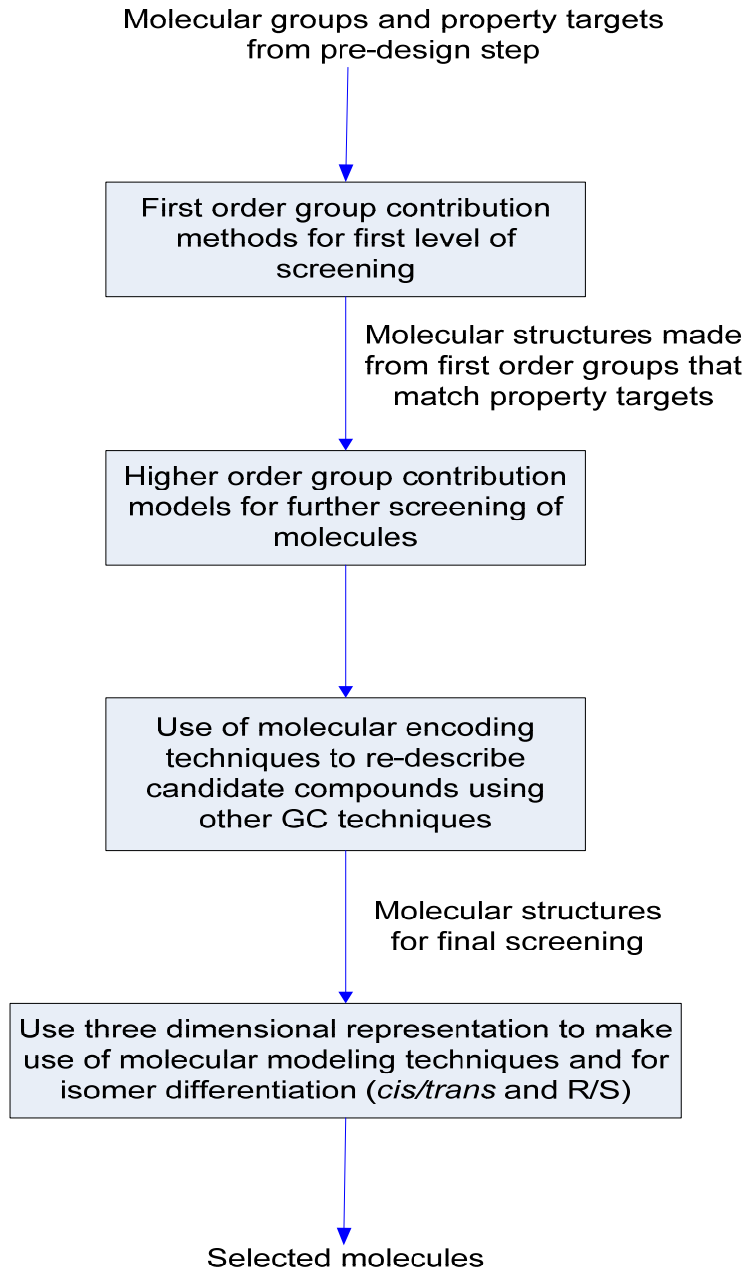


Figure 3.2: Multilevel Approach for Product Design

3.2. Computer Aided Molecular Design Techniques

Among the different techniques for physical property prediction, very few can be applied in the computer aided molecular design techniques mentioned in this dissertation. Most computer aided product/molecular design techniques utilize group contribution models (Achenie *et al.*, 2003). Topological index based QSAR/QSPR expressions have also found great applications in reverse problem formulations (Raman & Maranas, 1998; Camarda & Maranas, 1999) even though the available techniques to apply QSAR/QSPR relationships are limited. More detailed discussions of these methods have been presented in the previous chapter.

The purpose of CAMD algorithms is to solve the different classes of problems mentioned in chapter 1. In the first type, where only a database search is involved, search engines commonly look for a subset of the actual collection of molecules that satisfy the property constraints and molecular type constraints (if any) (Cabezas, 2000). However, for all the types of molecular design problems that include a generation step, the algorithms are required to solve them efficiently. The basic concept behind such algorithms is to generate molecular structures from the molecular fragments that satisfy a set of property constraints while obeying the feasibility requirements for the existence of a molecule. Even though a number of algorithms are available to solve CAMD problems, they can be broadly classified into three groups (Achenie *et al.*, 2003):

1. Mathematical programming – Solving an optimization problem as discussed in section 2.2 of chapter 2.
2. Stochastic optimization – The mathematical representation of the problem is solved using numerical stochastic methods like genetic algorithm.

3. Enumeration techniques – A combined mathematical and qualitative problem is solved by hybrid solution approaches.

3.3. Property Models

The input to a property model usually includes information such as composition and process conditions like temperature, pressure etc. and the outputs are the calculated/estimated property values. For the effective use of property models, the following features are identified (Gani & O'Connell, 2001):

1. The process conditions should not depend on the size of the system.
2. Property models are tools, which can provide the properties, which cannot be measured directly, from the quantities that can be measured directly.
3. Some properties predicted from the model may be relevant only at the specific process conditions used to predict its value.
4. Property models must have good extrapolating abilities.
5. The secondary variables like operating conditions, controlled variables, energy consumption and environmental impact must be subsets of the whole set of conditions and properties.
6. Because of the ability of computers to handle the computational load expected from complicated models, the actual form of the property models will be a parameterized mathematical formulation converted to a computer code.

Gani and O'Connell (2001) have suggested three distinct roles for property models. The first one is a service role, where the property model is to provide the

property values corresponding to the process conditions. This role is used primarily in simulators and the most important qualities expected from the models are accuracy and generality. For instance, during the simulation of a distillation column, the property models can be used to provide the values of fugacity equilibrium constants, vapor and liquid enthalpies, etc. when requested. One significant difficulty in this role is that, the property models have to be appropriately selected for getting the desired result. The second role is a service/advice role where the model provides information regarding the steps to be taken for the effective solution of the process simulation/design problem in addition to generating the property values. This role finds its application in the process design and synthesis problems. In synthesis problems, the solution is obtained in two steps. In the first step, the possible candidate formulations and/or process conditions (service and advice) are suggested. In the second step, the candidate formulations/conditions are verified (service role) and the solutions satisfying the requirements are selected. The most comprehensive role of a property model is the service/advice/solve role. This approach is typically used in integrated designs. Here, the identified property targets will serve as tools to connect the simulation and design problems. Therefore, the property values are identified by solving the simulation problem corresponding to the optimum process performance. The identified property values will form the design targets for the design problem to identify the process conditions that match the target (Gani, 2004b). In this way, the property models are decoupled from the process model because the property model is not needed in the simulation steps or in the design stage (Eden *et al.*, 2004; Eljack *et al.*, 2005).

3.4. Reverse Problem Formulation

In most process/product design problems, the computational complexity can be attributed to the constitutive equations because they are generally highly nonlinear. Eden *et al.* (2002) have shown that the reverse problem formulation can be successfully applied to process/product design problems to avoid the use of constitutive equations in the design because the targets for the design problem are functions of properties.

According to the method developed by Eden *et al.* (2004), the input to the process design problem is the desired process performance and the input to the molecular design problem are the molecular building blocks and the property targets identified in the process design step. The output of this algorithm will be the property values corresponding to the optimum process performance and the molecular structures corresponding to the property targets identified in the process design step. The advantage of this approach is that, the designer is not committing to any specific components during the design. This methodology is illustrated in figure 3.3. One of the challenges in following such an algorithm is that, the process design problem is solved in terms of the properties and not in terms of components. Unlike mass and energy, properties are not conserved, however, there is a way to systematically track properties, which will be explained in the next chapter.

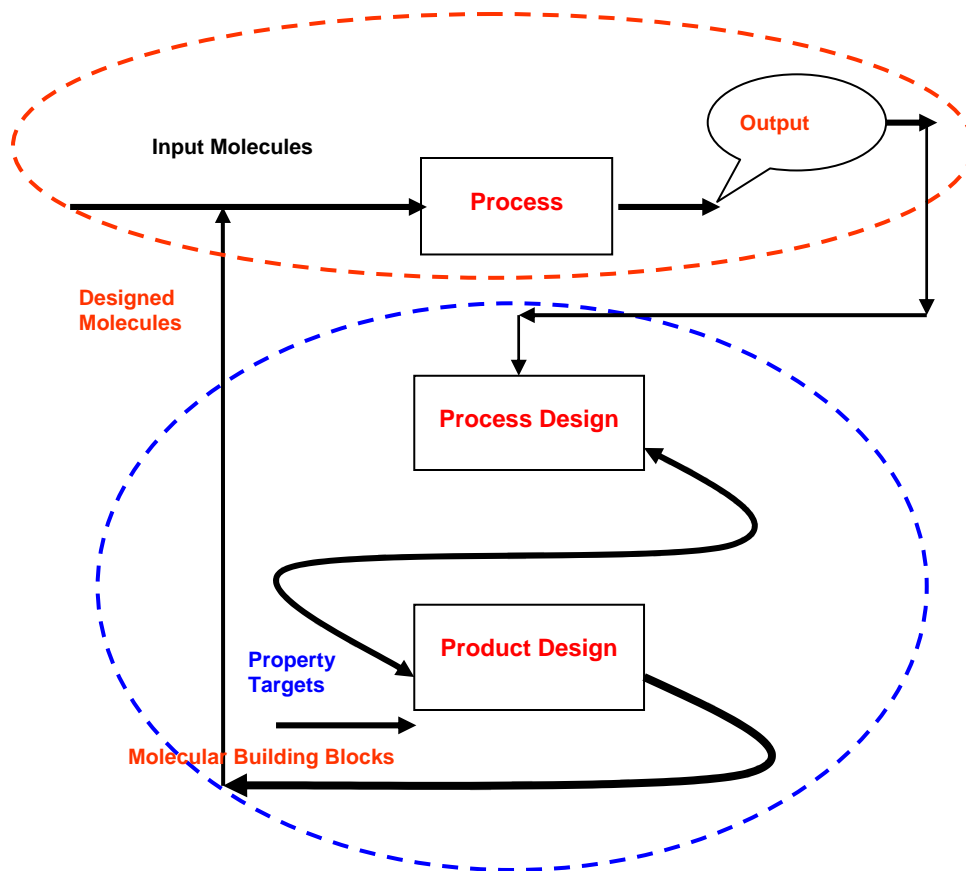


Figure 3.3: Simultaneous Consideration of Process and Product Design

3.5. Reverse Problem Formulation Methodology

The procedure developed by Eden *et al.* (2002) for decoupling the constitutive equations is illustrated in figure 3.4. The result will be two reverse problems. The first reverse problem is the reverse of a simulation problem. Here, the objective is to determine the process variables corresponding to the given input variables, equipment parameters and desired output parameters. The second reverse problem is the reverse of a property prediction problem, in which the molecular structures corresponding to the identified property targets are generated.

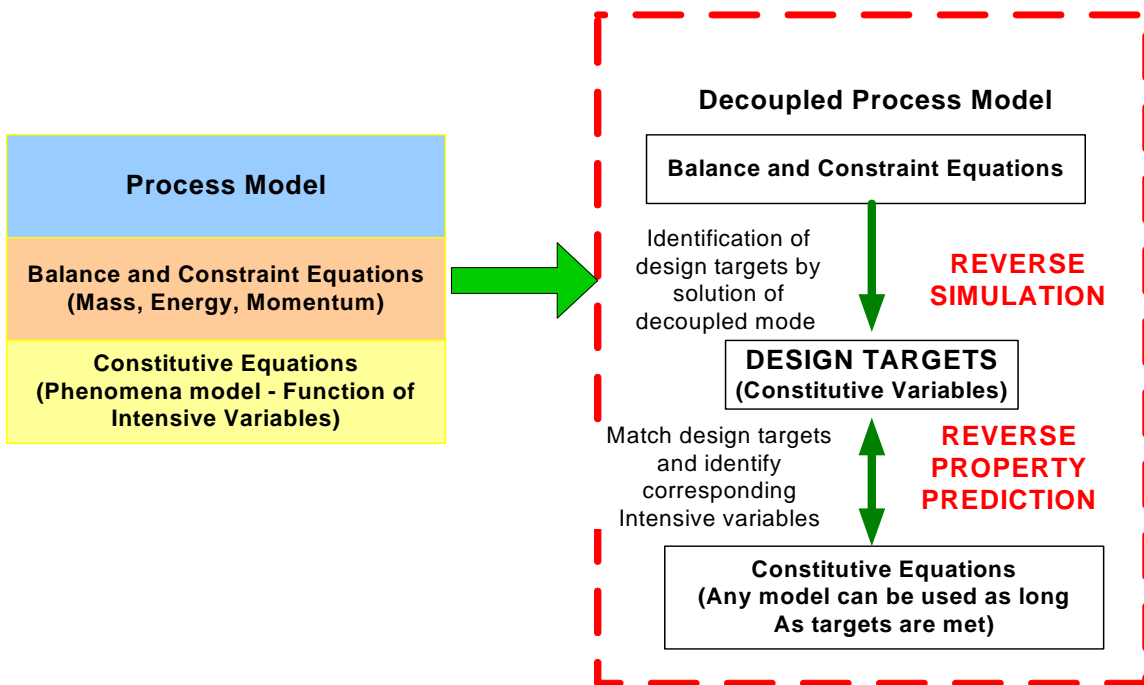


Figure 3.4: Reverse Problem Formulation (adapted from Eden *et al*, 2002)

As the complex constitutive equations are eliminated from the system of equations to be solved for the targets, the solution step is easy. In addition, any number of property models can be used for the second reverse problem as long as the target constitutive variable values are matched. It is possible to have more than one solution since the algorithm involves a matching procedure. Therefore, a performance index can be defined and evaluated for all identified solutions to determine the optimal solution.

3.6. Summary

This chapter provides a brief overview of the field of computer aided molecular design. Because of the huge amount of data involved and the non-linear nature of the mathematical formulations involved in process and product design problems, computer aided solution techniques provide convenient ways to reach solutions. A brief overview

of the classifications of the different types of approaches to solving CAMD problems has been presented. The current techniques are useful in solving different types of problems, however, the computational expenses for using these techniques are very high and the global optimality of the solutions cannot be ensured in many cases.

The different roles of property models have been analyzed. The three different roles of property models are described and the concept of reverse problem formulation has been explained to illustrate the advantages of applying RPF in product design. Finally, the application of RPF in the simultaneous consideration of process and product design problems has been introduced along with a targeting method to decouple the property models from design equations.

The traditional approach followed in computer aided molecular design problems is the multilevel approach. However, the iterative nature of this method makes it cumbersome and less efficient. It is also possible that many of the potential solutions will not be screened in the initial stages. Therefore, any new algorithms in the field of computer aided molecular design should focus on providing non-iterative solution strategies.

4. Integrated Process and Molecular Design

The objective of this dissertation is to address the simultaneous consideration of process and product design through development of a systematic non-iterative procedure to approach design from a property perspective. As mentioned before, in order to follow the targeting approach for integrated process and product design, there should be a common property based platform. The property clustering technique introduced by Shelley and El-Halwagi (2000) provides the tools to track properties. In addition, there should be tools that make use of the available property models to solve reverse problems. This chapter presents the techniques developed in this dissertation research to solve the integrated process and product design problems from a property perspective.

4.1. Property Clustering Techniques

4.1.1. Property Operator and Cluster Formulation

There are many processes where the basis for design is not the actual chemical components due to the non-homogeneous nature of the process streams and the multitude of chemical components involved. Instead, the designer focuses on the properties that drive the process. One example of a property driven problem is the design of paper with a specified quality. Since the basic component of all types of paper is cellulose, the quality cannot be defined in terms of components and/or composition. Instead, the quality is specified in terms of the physical properties (Eden *et al.*, 2004). However, the main

limitation for designing a process based on properties is that, unlike mass and energy, properties are not conserved. The concept of property clustering has been introduced to resolve this limitation by mapping the property relationships into a low dimensional domain (Shelley & El-Halwagi, 2000). The property clusters are formed based on property operators, which are functions of actual physical properties that obey linear additive rules (Shelley & El-Halwagi, 2000; Eden *et al.*, 2004). Therefore, the first step in a property based design algorithm is to find the ideal property operator corresponding to the non-linear properties. So, for a mixture made up of N_s streams and described by j properties, the property operator, $\Psi_j(P_{jM})$ corresponding to property P is formulated as follows:

$$\Psi_j(P_{jM}) = \sum_{s=1}^{N_g} \frac{F_s}{\sum_{s=1}^{N_s} F_s} \Psi_j(P_{js}) = \sum_{s=1}^{N_g} x_s \Psi_j(P_{js}) \quad (4.1)$$

where, $\Psi_j(P_{js})$ is the operator of the j^{th} property P_{js} of stream s .

It can be seen that the property operators obey linear mixing rules irrespective of the nature of the actual property. One classic example is density. The mixing rules for density are not linear. However, the property operator defined for density obeys linear mixing rules as shown in eq. (4.2):

$$\frac{1}{\rho_M} = \sum_{s=1}^{N_g} x_s \frac{1}{\rho_s} \quad \Psi_j(P_{jM}) = \frac{1}{\rho_M} \quad \Psi_j(P_{js}) = \frac{1}{\rho_s} \quad (4.2)$$

In a process system, the properties may be of different units and magnitudes. In order to make the properties dimensionless and of similar magnitude, the property operators are divided by appropriately chosen reference operators. The normalized property operator thus obtained is defined as:

$$\Omega_{js} = \frac{\psi_j(P_{js})}{\psi_j(P_j^{ref})} \quad (4.3)$$

The Augmented Property Index, AUP is defined as the sum of all the dimensionless property operators present in the system:

$$AUP_s = \sum_{j=1}^{NP} \Omega_{js} \quad (4.4)$$

Finally, the property cluster C_{js} for property j is defined as:

$$C_{js} = \frac{\Omega_{js}}{AUP_s} \quad (4.5)$$

The property cluster of one property can be understood as the fraction of one property in the whole system of properties.

4.1.2. Conservation Rules

The formulation of property clusters ensures that they obey two fundamental conservation rules, i.e. intra-stream conservation and inter-stream conservation. Intra-stream conservation implies that, all the clusters corresponding to the properties in each stream s add up to unity. Therefore, for a system of NP properties, if the cluster values of $(NP-1)$ properties are known, the NP^{th} property is intrinsically given. Inter-stream conservation implies that, after mixing different streams, the cluster values can be calculated as the weighted average of the contributions from their individual flowrates. Therefore, when two streams S_1 and S_2 are mixed, the straight line connecting those two streams will provide the locus of all cluster values for all possible mixture combinations. The conservation rules for clusters are given in eqs. (4.6) and (4.7). If the number of properties is limited to three, the clusters can be represented on a ternary diagram and representations of the corresponding conservation properties of the clusters are shown in figure 4.1 and 4.2 (Eden *et al.*, 2004), respectively:

$$\sum_{j=1}^{N_C} C_{js} = 1 \quad (4.6)$$

$$C_{jM} = \sum_{j=1}^{N_C} \beta_s \cdot C_{js} \quad (4.7)$$

where N_C is the total number of clusters and β_s is the mixing ratios of the clusters.

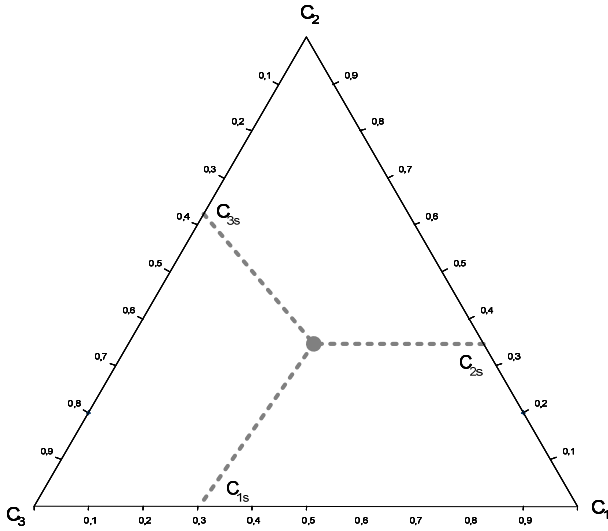


Figure 4.1: Visualization of Intra-stream Conservation of Clusters

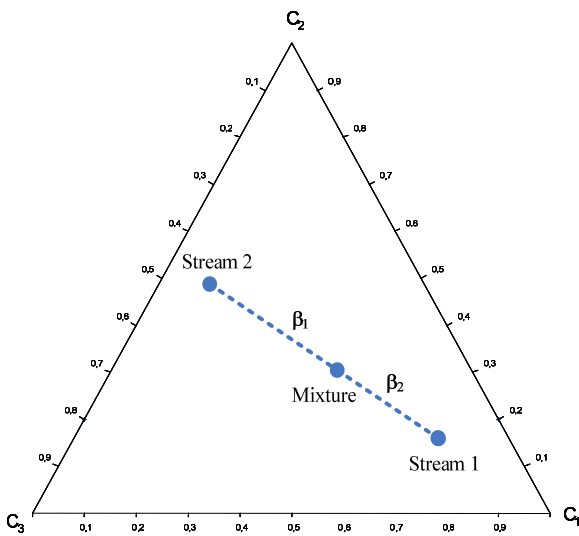


Figure 4.2: Visualization of Inter-stream Conservation of Clusters

As the cluster values are conserved after mixing the streams, the cluster value for property j in the mixture can be formulated as follows:

$$C_{jM} = \frac{\Omega_{jM}}{AUP_M} \quad (4.8)$$

Similarly, the expressions for normalized property operator, cluster arm, β_s and AUP_M can be obtained as follows (Eden *et al.*, 2004):

$$\Omega_{jM} = \sum_{s=1}^{N_s} x_s \cdot \Omega_{js} \quad (4.9)$$

$$\beta_s = \frac{x_s \cdot AUP_s}{AUP_M} \quad (4.10)$$

$$AUP_M = \sum_{s=1}^{N_s} x_s \cdot AUP_s \quad (4.11)$$

4.1.3. Visualization Techniques

As long as the number of properties is less than or equal to three, they can be represented on a ternary diagram and the property change due to mixing of streams can be tracked visually using the conservation rules. After obtaining the cluster values, they can be plotted inside a ternary diagram. Since the clusters represent virtual properties, this technique provides a unique way for tracking the actual properties.

The above-mentioned methodology can be used to estimate whether the recycle and mixing of streams can provide the required performance. Suppose, the sink region (units capable of processing the sources) is a hexagon as shown in figure 4.3. The

systematic procedure for the identification of the sink region is given in section 4.1.4. As mentioned before, all the possible cluster combinations of two streams will lie on the straight line connecting those two points. Therefore, if the straight line passes through the sink region, the two streams can be mixed to get the required output, S . In figure 4.3, S_1 and S_2 can be mixed to get the optimum output where as neither of these streams can be mixed with S_3 because, no combination with S_3 will pass through the sink region.

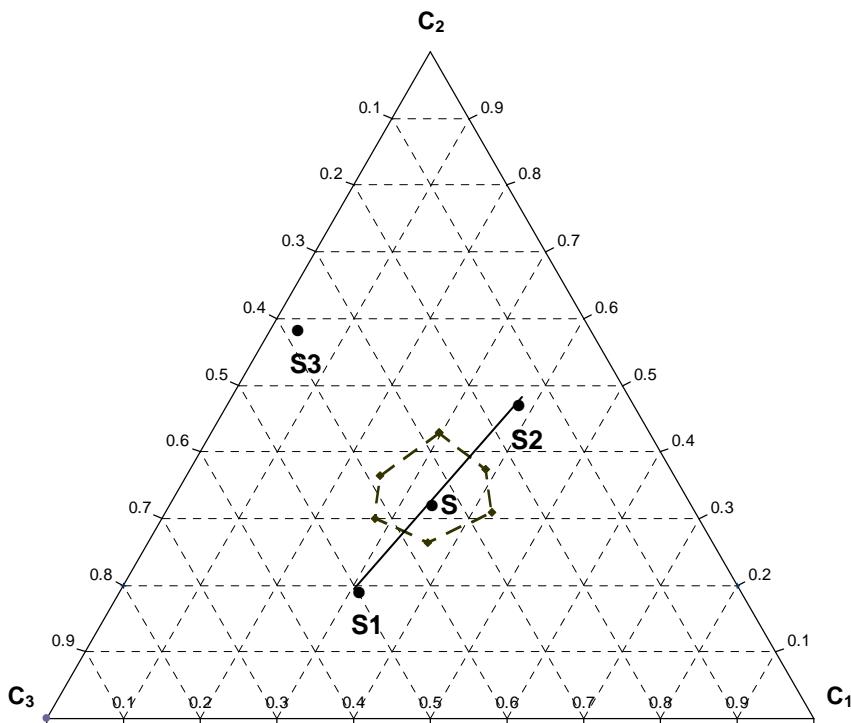


Figure 4.3: Mixing of Streams

It is to be noted that the cluster values represent the possible proportions of properties. Therefore, the cluster value inside the sink region alone will not ensure that the properties are in the correct range. In order to make sure that the properties match the sink requirements, the AUP values of the sink and source streams must also match. In

addition, the sink may have upper and lower limits of capacity for its proper functioning. Therefore, the flowrate of sources or mixture of sources must be within those limits.

The conditions described above are all necessary conditions and if any of those are not satisfied, the individual flowrates or feed compositions must be changed in order to be accepted by the sink. Addition of a new source is possible as long as all conditions are satisfied.

4.1.4. Identification of Feasibility Region

The actual identification of the feasibility region of a sink is a tedious procedure because it requires a one to one mapping of an infinite number of feasible points. El-Halwagi *et al.* (2004) developed a method for mapping the feasibility region without enumeration of this infinite number of feasible points. In the first step, the feasibility region is overestimated by plotting only the minimum and maximum values of clusters and connecting them using straight lines. This step reduces the search space significantly.

Consider a sink with three targeted properties. Suppose, each property is bounded by a lower and upper limit. Therefore, using the property operators and clustering techniques, the following equations are developed:

$$P_{j,\text{sink}}^{\min} \leq P_j \leq P_{j,\text{sink}}^{\max} \quad , \quad \Omega_{j,\text{sink}}^{\min} \leq \Omega_j \leq \Omega_{j,\text{sink}}^{\max} \quad (4.12)$$

$$C_{1,\text{sink}}^{\min} = \frac{\Omega_{1,\text{sink}}^{\min}}{\Omega_{1,\text{sink}}^{\min} + \Omega_{2,\text{sink}}^{\max} + \Omega_{3,\text{sink}}^{\max}} \quad (4.13)$$

$$C_{1,\text{sink}}^{\max} = \frac{\Omega_{1,\text{sink}}^{\max}}{\Omega_{1,\text{sink}}^{\max} + \Omega_{2,\text{sink}}^{\min} + \Omega_{3,\text{sink}}^{\min}} \quad (4.14)$$

$$C_{2,\text{sink}}^{\min} = \frac{\Omega_{2,\text{sink}}^{\min}}{\Omega_{1,\text{sink}}^{\max} + \Omega_{2,\text{sink}}^{\min} + \Omega_{3,\text{sink}}^{\max}} \quad (4.15)$$

$$C_{2,\text{sink}}^{\max} = \frac{\Omega_{2,\text{sink}}^{\max}}{\Omega_{1,\text{sink}}^{\min} + \Omega_{2,\text{sink}}^{\max} + \Omega_{3,\text{sink}}^{\min}} \quad (4.16)$$

$$C_{3,\text{sink}}^{\min} = \frac{\Omega_{3,\text{sink}}^{\min}}{\Omega_{1,\text{sink}}^{\max} + \Omega_{2,\text{sink}}^{\max} + \Omega_{3,\text{sink}}^{\min}} \quad (4.17)$$

$$C_{3,\text{sink}}^{\max} = \frac{\Omega_{3,\text{sink}}^{\max}}{\Omega_{1,\text{sink}}^{\min} + \Omega_{2,\text{sink}}^{\min} + \Omega_{3,\text{sink}}^{\max}} \quad (4.18)$$

The overestimation will not provide the actual feasibility region. However, it can be ensured that no points outside the overestimated region can be a solution. The equations provide a point on each line segment joining the overestimated feasibility region boundaries. Since these points are also part of the true feasibility region, the line segments joining these points represent different possible combinations of these points. Therefore, the hexagon obtained after connecting these points will form the boundaries of feasibility region.

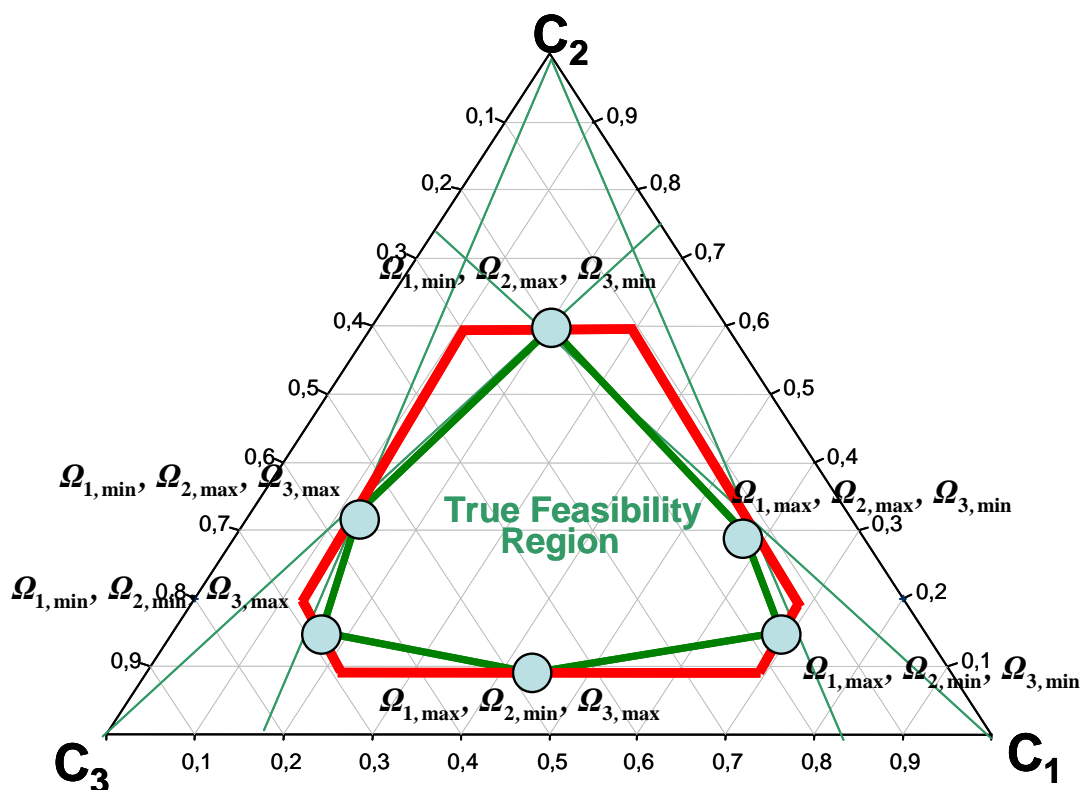


Figure 4.4: Feasibility Region on a Ternary Diagram (Eden *et al.*, 2004)

4.2. Molecular Property Operators and Clusters

The property clustering techniques presented in the previous section have been extended to different areas of process and product design. To introduce them into molecular design, the group contribution methods (GCM) have been used. For that, an interesting similarity between the formation of property operators and the property function models in GCM can be utilized as long as the molecule is described with only with first order groups.

In GCM, the property function for one particular property of any molecule is calculated as the sum of the property contributions of the individual molecular fragments

(Constantinou & Gani, 1994) whereas in property clustering, the property operators of each fraction are added up to give the property operator of one particular property.

So it is possible to employ a similar treatment to GCM to convert it into a powerful tool for molecular design in the cluster space as explained below (Eljack *et al.*, 2007b; Kazantzi *et al.*, 2007)

If P_{jg} is the contribution of property j from group g and n_g is the total number of that group in the molecule, then the molecular property operator, ψ_j can be defined as:

$$\psi_j(P_j) = \sum_{g=1}^{N_g} n_g P_{jg} \quad (4.19)$$

The number of properties that can be predicted using group contribution methods is limited. However, many empirical and non-empirical expressions exist that relate the group contribution properties to some of the non-GC properties. If such an expression exists, then, the property target can be obtained in terms of the GC property for a given non-GC property to solve the design problem.

As in the case of property operators, the complexity and non-linearity in the actual group contribution relationship is masked inside the relationship between the molecular property operator and the groups. The operators obey simple linear mixing rules (Eljack *et al.*, 2007b; Eljack & Eden, 2008).

Following the same procedures and logic used to develop the original property clusters, it is possible to define the normalized molecular property operator Ω_j^M , Augmented Property Index AUP^M , and the molecular property cluster C_j^M .

$$\Omega_j^M = \frac{\psi_j^M(P_{ji})}{\psi_j^{ref}(P_{ji})} \quad AUP^M = \sum_{j=1}^{N_p} \Omega_j^M \quad C_j^M = \frac{\Omega_j^M}{AUP^M} \quad (4.20)$$

Similar to the original property clusters, molecular property clusters also have two fundamental properties, intra- and inter-molecular conservation. Similar to the intra-stream conservation rule for processes, the intra-molecular conservation implies that, the all the molecular clusters corresponding to one property in a molecule must sum to unity as shown in eq. (4.21):

$$\sum_{j=1}^{N_c} C_j^M = 1 \quad (4.21)$$

Inter-molecular conservation implies that after “mixing” different molecular groups, the individual cluster values will be conserved. That means, the cluster value of the molecular string will be estimated from the weighted average of clusters corresponding to the number of groups in each operator:

$$\Omega_{j,mix}^M = \sum_{g=1}^{N_g} n_g \Omega_{jg} \quad (4.22)$$

$$C_{j,mix}^M = \sum_{g=1}^{N_g} \beta_g C_{jg} \quad (4.23)$$

The proofs for the above expressions are similar to the property cluster equations and have been presented by Eljack *et al.* (2007).

4.3. Visual Solution of Molecular Design Problem

The algorithm for solving a molecular design problem using the molecular property operators is similar to the process design problem using property operators. The step-wise procedure for converting the molecular property data into cluster space for a visual solution is given in table 4.1:

Table 4.1: Visual Molecular Design Algorithm

<i>Step</i>	<i>Description</i>	<i>Equations</i>
1	Find the molecular property operator from group contribution expressions corresponding to each property and obtain the group contribution corresponding to that property	-
2	Convert the property contribution into normalized molecular property operator.	(4.19)-(4.20)
3	Calculate <i>AUP</i> values	(4.20)
4	Calculate molecular cluster values corresponding to each group and plot those points on a ternary diagram.	(4.20)-(4.23)

As explained before, an analogous relation between the property clusters used in process design and molecular clusters permit a similar methodology for “mixing” or combining different molecular building groups. Different combinations of molecular groups can be tried based on the nature of the final product requirement. Similar to two different streams, combination of two molecular groups will form a straight line upon “mixing”. Now, to build meaningful and complete molecules that satisfy all the target

properties, the following rules are to be obeyed (Eljack *et al.*, 2007b; Eljack & Eden, 2008).

Rule 1: The visualization arm β describes the location of the new molecular fragment when two groups are combined in a ternary diagram

$$\beta_1 = \frac{n_1 \cdot AUP_1}{n_1 \cdot AUP_1 + n_2 \cdot AUP_2} \quad (4.24)$$

Rule 2: The Free Bond Number (*FBN*) is the number of free bonds in each molecular string (Eljack *et al.*, 2007b, Eljack & Eden, 2008) and is represented mathematically in eq. (4.25):

$$FBN = \sum_{g=1}^{N_g} n_g FBN_g - 2 \left(\sum_{g=1}^{N_g} n_g - 1 \right) - 2N_r \quad (4.25)$$

Where, N_r is the number of rings in the final molecule and FBN_g is the number of free bonds in each group. Again, the constraint here is the molecule's *FBN* should be zero. This is to ensure a complete molecular structure with no charge/no free bonds in the final molecule.

Rule 3: The location of the final molecule depends only on the type and numbers of each group forming the molecule. It is independent of the order of mixing.

The proof for this rule can be obtained from Rule 1, which provides an expression for the visualization arm. Therefore, it is clear from its definition that, the relative positions of the molecular groups are only functions of the number of occurrences of each group and their *AUP*. As *AUP* is only a function of the property contribution of the specific group, the order sequence of mixing of the groups will not be a factor in the final position of a molecular fragment.

Rule 4: The cluster location of the designed molecule should fall inside the feasibility region of the sink identified through the algorithm explained in the previous section.

Rule 5: The *AUP* value of the designed molecule must be within the *AUP* range of the sink.

Rule 6: The three necessary conditions for the designed molecule to match the target properties are: the cluster location of the final molecule must be in the feasibility region identified for the sink, the *FBN* of the final formulation is zero, and the *AUP* has to be within the limit of the sink, If any of these conditions are violated, the designed molecule will not be feasible. The sufficient condition is the matching of the actual physical properties. If a combination of molecular groups satisfy all the necessary conditions, their property values must be back calculated using the group contribution methods to ensure that they are valid formulations.

In the example shown in figure 4.5, the mixing/combination of three molecular groups, G_1 , G_2 , G_3 to produce a molecule M is shown. If, M satisfies the *AUP* and *FBN* constraints, it will be a feasible molecule for further study.

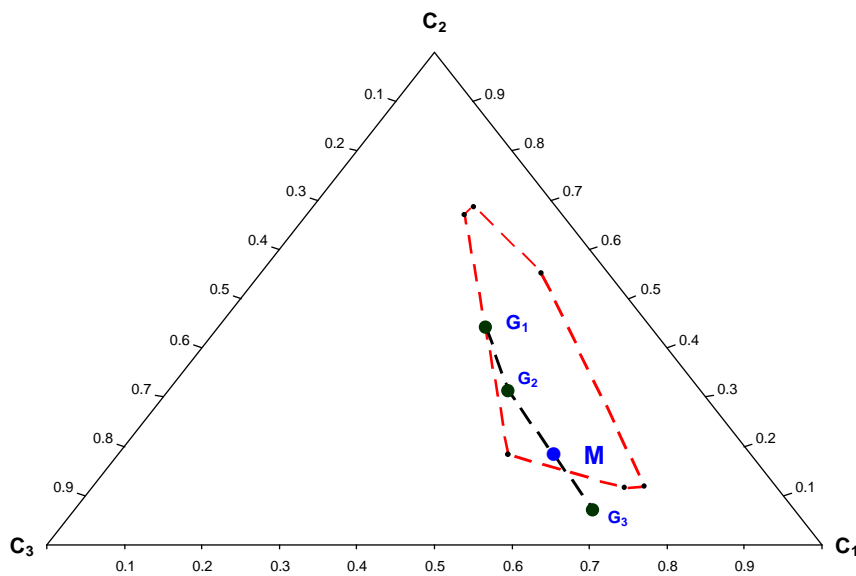


Figure 4.5: Mixing of Molecular Groups

The clustering concept enables the consideration of any number of components and streams in the process design and any number of molecular groups in the product design step. The dimensionality of the problem is affected only by the number of properties of interest and if the number of properties is limited to three, a visual approach can be used whereas an algebraic approach must be applied if more properties are required to adequately describe the system.

As the product design requires the input in terms of GC properties, the process design step must provide the output in terms of GC properties corresponding to the optimum process performance. After identifying the target properties, the input to the product design will consist of the property targets along with the molecular groups to be

considered. If the identified property targets are GC properties, the GC models can be used directly in the product design step. If not, the property targets have to be redefined in terms of GC properties using empirical relationships between the desired properties and GC properties. The property targets will form the feasibility region and the molecular groups will form discrete points on the ternary diagram. The molecular synthesis will be carried out as described in the previous section and as the target properties corresponding to the optimum process performance, the molecules designed will match the performance targets. A visual representation of the simultaneous approach to process and product design is shown in figure 4.6.

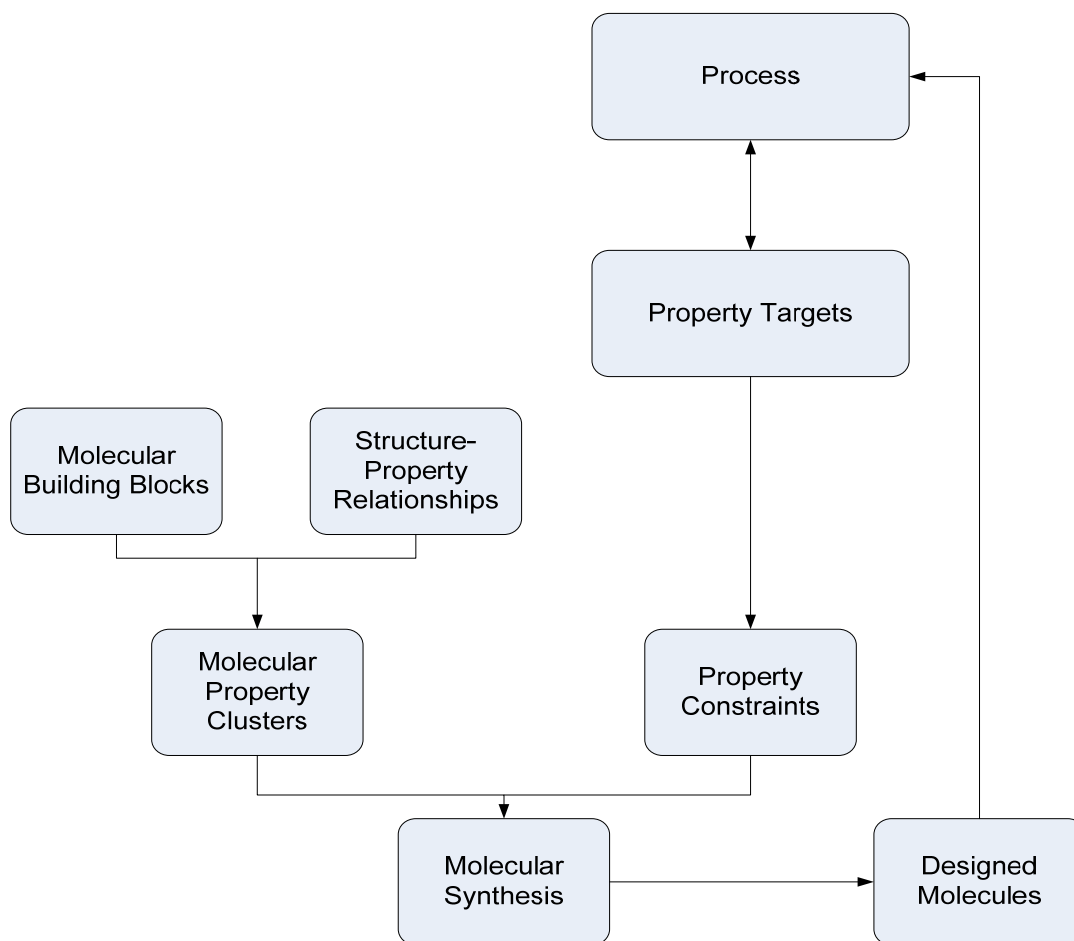


Figure 4.6: Simultaneous Process and Product Design Framework

4.4. Limitations of the Visual Approach in Molecular Design

The molecular property operators developed in the above-mentioned method can only be used for the design of simple monofunctional molecules as they are based only on first order groups (Kehiaian, 1983; Wu & Sandler, 1989, 1991; Marrero & Gani, 2001). The first order group contribution method has limited accuracy especially when dealing with polyfunctional molecules and cyclic molecules. In addition, first order groups cannot capture proximity effects or differentiate between isomers (Kehiaian,

1983; Wu & Sandler, 1989, 1991; Marrero & Gani, 2001). To overcome these limitations of the first order groups, higher order group effects must be incorporated into the design.

The applicability of the visual approach is limited to problems that can be adequately described using three properties. However, there are many systems that require more than three properties for sufficient representation. In the molecular design stage, all possible combinations of different molecular groups are to be analyzed, in order to get a complete solution of possible compounds. This will be a tedious process if there are many first order groups, because for n groups the total number of combinations will be ${}^n C_n + {}^n C_{n-1} + {}^n C_{n-2} + \dots + {}^n C_2$. Another serious limitation of the visual approach is that, even though it may help rule out some infeasible combinations in simple designs, the required numbers of molecular groups is mainly a function of their *AUP* values. Therefore, the relative positions in the ternary diagram will not always provide the visual insights expected from it. In addition, the relative position of the clusters being included in the feasibility region is only one of the requirements for a feasible molecule. It has to satisfy the *AUP* constraint as well as described earlier. Therefore, for each combination, the *AUP* constraint has to be verified separately in this approach, which again makes the procedure tedious for large problems.

There are limitations associated with the property models as well. Even though the GC models predict the properties of molecules with reasonable accuracy, there are occasions when one or more of the molecular groups of interest are not available in the literature. Similarly, even if the molecular group is available, the property contribution corresponding to that group may not be available. It is to be noted that, if any of the property information of one or more of the groups is not available, the design cannot be

completed. One option available at this point is to regress the contribution of that molecular group. However, this may be a lengthy procedure and may not always be practical especially when the property values are not available (Gani *et al.*, 2005). One of the feasible approaches is to make use of the property models from the combined group contribution-connectivity index method. Therefore, the algorithm for molecular design has to be modified to include GCM+CI models in the cluster space.

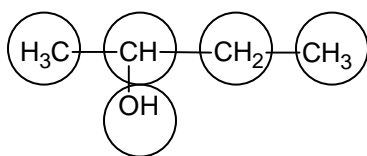
Many of the properties encountered in design may require more structural information than what is provided by group contribution techniques. Especially when dealing with complex molecules, the structural details become more relevant for the determination of properties. The recent developments in the field of QSAR and QSPR studies provide many relationships based on molecular structures to predict properties. However, the inverse problem formulations with these relationships form non-linear equations and their solutions require a lot of computational effort and often lead to degenerate solutions. The molecular signatures described in the previous section can be used as a common tool to translate properties back to molecular structures and an algorithm is needed for such a systematic enumeration.

4.5. An Algebraic Approach for Molecular Synthesis with Higher Order Groups

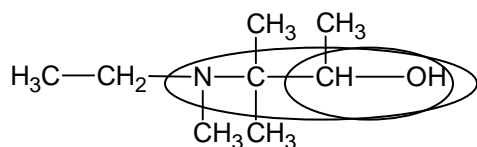
As discussed in the previous section, one of the main drawbacks of the current visual solution method for molecular design problems is its inability to consider higher order molecular groups. In the new algorithm developed in this dissertation research, an algebraic method has been used to include the contributions of higher order molecular groups identified by Marrero and Gani (2001) in the reverse problem formulation.

The algebraic approach for property integration through componentless design of processes has been already developed (Qin *et al.*, 2004). Nevertheless, this approach was limited to the design of simple molecules comprised of two molecular building blocks even though the approach was helpful to solve for any number of properties. In addition, the applicability of this method is limited to the design of non-cyclical compounds. A similar approach was followed by Eljack *et al.* (2007a) for molecular design with first order groups by taking advantage of the analogous nature of first order molecular property operators to the traditional property operators. Presented here is a generalized algebraic approach for designing molecules with any number of first order groups including the possible contributions from second and third order groups.

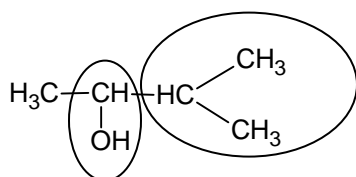
To include the effects of second and third order groups in the property prediction, it is possible to utilize the linear additive rules of higher order groups. Further, both higher order groups have first order groups as their building blocks (Marrero & Gani, 2001) and hence they can be considered as combinations of different first order groups. It should be noted that any kind of overlapping of molecular fragments in different higher order groups is permitted. That means, one first order group can be a part of more than one higher order group since the higher order groups represent different kinds of interactions among the molecular fragments. However, it must be ensured that no group is completely overlapped by another group. If the building blocks selected for constructing a molecule can generate such second order groups or third order groups, the group with more first order groups is to be selected to form the second order or third order group to avoid redundant description of the same molecular fragment (Marrero & Gani, 2001). Figure 4.7 describes the different scenarios of group overlapping:



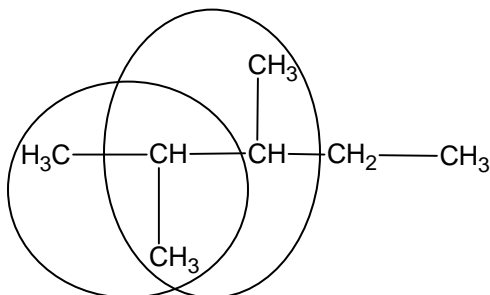
First order groups



Second order groups with complete overlapping



Second order groups without overlapping



Second order groups with partial overlappings

Figure 4.7: Second Order Group Formation

In figure 4.7, a molecule's building blocks in the form of first order groups is shown. In the second figure, the second order groups are NCCH_2OH and CH_2OH . Here only the contribution from the former group should be considered because the group CH_2OH is completely overlapped by NCCH_2OH . In the third example, there are two second order groups: CH_2OH and $(\text{CH}_3)_2\text{CH}$ with no overlapping. In the last example,

there are two second order groups, $(\text{CH}_3)_2\text{CH}$ and $(\text{CH}_3)\text{CH}-(\text{CH}_3)\text{CH}$. Here, two distinct second order groups share one CH group and one CH_3 group. However, since they represent two different types of interactions, the contributions from both these second order groups have to be considered when applying the property model.

4.5.1. General Problem Statement

Generate all possible molecules that can be built from N_g molecular groups with N_j target properties. There is a set of constraints for each property, which can be represented as the following:

$$P_{ij}^{lower} \leq P_{ij} \leq P_{ij}^{upper} \quad j= 1, 2, \dots, N_j; \quad i=1, 2 \dots \quad (4.26)$$

where i is the index of molecules and j is the index of properties.

4.5.2. Algebraic Approach for Solving the Molecular Design Problem

It can be seen from eq. (4.26) that, each property range can be expressed as two inequality expressions (Qin *et al.*, 2004): one for the lower constraint and one for the upper constraint. This range can be estimated as a function of molecular groups from GCM expressions (Eljack *et al.*, 2007a). This facilitates a matching of the property targets with molecular constraints. Equation (4.26) can be rewritten in terms of normalized property operators as:

$$\Omega_j^{\min} \leq \Omega_{ij} \leq \Omega_j^{\max} \quad (4.27)$$

Here, Ω_{ij} is the normalized property operator of molecule i . To estimate its value, first calculate the normalized property operator based on first order estimation, Ω_{ijf} as

$$\Omega_{ijf} = \sum_{g=1}^{N_g} n_g \Omega_{jg1} \quad (4.28)$$

where, Ω_{jg1} is the normalized property operator of first order group, g . The next step is to estimate the contributions of any second order groups in the molecule. The following rules must be followed at this stage:

Rule 7: Second order groups have first order groups as building blocks.

Rule 8: Second order groups can only be formed from complete molecular fragments. For instance, to form the second order group CH-(CH₃)-CH-(CH₃), there must be two -CH- and two (CH₃) groups. It is not possible to consider a CH-(CH₃) group as a half second order group.

Rule 9: If a second order group completely overlaps another second order group, then only the larger of the two groups is chosen in order to prevent the redundant description of the same molecular fragment.

Let $(k: n)$ be the set of first order groups that are the building blocks of one second order group s and $(n_{gk}:n_{gn})$ is the set of occurrences of those groups present in the molecule. If η is the number of occurrences of one particular first order group in a

selected second order group, and n_{gs} is the number of second order groups which can be generated from those first order groups, then:

$$n_{gs} = \text{Min} \left(\frac{n_{gk}}{\eta_k} : \frac{n_{gn}}{\eta_n} \right) \quad (4.29)$$

Here, n_{gk}/η_k is the contribution of first order group k to the second order group, n_{gn}/η_n is the contribution of the first order group n to the second order group such that eq. (4.29) will give the potential contributions from each first order group to the specified second order group:

$$\left(\frac{n_{gk}}{\eta_k} : \frac{n_{gn}}{\eta_n} \right) \quad (4.30)$$

According to Rule 8, the lowest of those numbers will give the number of that second order group. For instance, if there are three CH₃- groups and two -CH- groups in the molecule, then there will be one (CH₃)₂CH second order group. If Ω_{jg2} is the property contribution from the second order groups, then the normalized property operator for the property contributions from second order groups, Ω_{ijs} can be calculated as follows:

$$\Omega_{ijs} = \sum_{s=1}^{N_s} n_{gs} \Omega_{jg2} \quad (4.31)$$

n_{gs} must be rounded down to the nearest integer value before applying it in eq. (4.31) because the number of second order groups cannot be a fractional number. The above equation can predict the property contribution from second order groups in most molecules. However, in some rare occasions when some of the second order groups are completely overlapped by some bigger second order groups and some of the former groups are not overlapped, the equation will not account for the contribution from the unoverlapped group. For instance, if there are two –CH-, two OH- and one CN- groups in the molecular structure, the contributions from the CHOH group and the CNCHOH group must be incorporated as one CHOH group that is not overlapped by a CNCHOH group. To treat such a situation algebraically, consider that $(n_{gk}:n_{gn})$ has subsets of smaller second order groups $(n_{gl}:n_{gm})$ with some of the first order components of $(n_{gk}:n_{gn})$ and n_{gs}^* is the number of the second order groups which are not overlapped, then:

$$n_{gs}^* = \left[\text{Min} \left(\frac{n_{gl}}{\eta_l} : \frac{n_{gm}}{\eta_m} \right) - \text{Min} \left(\frac{n_{gk}}{\eta_k} : \frac{n_{gn}}{\eta_n} \right) \right] \quad (4.32)$$

According to Rule 8, n_{gs}^* must be rounded down to the nearest integer. If Ω_{jg2}^* is the contribution from the unoverlapped smaller second order groups, then the normalized property operator for the property contributions from the smaller second order groups, Ω_{ijs}^* can be calculated as:

$$\Omega_{ijs}^* = \sum_{s=1}^{N_s} n_{gs}^* \Omega_{jg2}^* \quad (4.33)$$

The third order groups have been identified by following the same criteria used for second order groups, but for a different class of compounds (Marrero and Gani, 2001). Therefore, the rules applied for forming second order molecular operators should be obeyed for generating third order groups as well. Therefore, following the same methodology for generating the second order molecular operators, the third order molecular operators have been formed as shown below:

$$\Omega_{ijt} = \sum_{t=1}^{Nt} n_{gt} \Omega_{jg^3} \quad (4.34)$$

$$\Omega_{ijt}^* = \sum_{t=1}^{Nt} n_{gt}^* \Omega_{jg^3}^* \quad (4.35)$$

Here, t is the index for third order groups. Here, n_{gt} and n_{gt}^* have been calculated with the groups corresponding to the third order group. Now, the normalized property operator for molecule i can be calculated as:

$$\Omega_{ij} = \Omega_{ijf} + \Omega_{ijs} + \Omega_{ijt} + \Omega_{ijs}^* + \Omega_{ijt} + \Omega_{ijt}^* \quad (4.36)$$

It is evident from eq. (4.27) that, for each property, there will be two inequality expressions in the cluster space – one for the minimum value and one for the maximum value (Qin *et al.*, 2004). Therefore, there will be $2N_p$ inequality expressions representing all the possible solutions. To solve for n_g , combine eqs. (4.27) and (4.36) and split them

into two equations for each property. Then, calculate the minimum and maximum values of *AUP* for the given property constraints.

$$\Omega_j^{\min} \leq \Omega_{ij} \quad \Omega_{ij} \leq \Omega_j^{\max} \quad (4.37)$$

For example, if there are four properties of interest, there will be eight inequality expressions from which eight subsets can be developed. For each subset, there will be four equations. These equations will provide all possible ways the properties can be combined without violating the property constraints. For the normalized molecular operators ($\Omega_1, \Omega_2, \Omega_3, \Omega_4$) the combinations that make up the eight subsets of equations are given below (Eljack *et al.*, 2007a):

$$\begin{array}{ll} (\Omega_1^{\max}, \Omega_2^{\min}, \Omega_3^{\min}, \Omega_4^{\min}) & (\Omega_1^{\min}, \Omega_2^{\max}, \Omega_3^{\max}, \Omega_4^{\max}) \\ (\Omega_1^{\min}, \Omega_2^{\max}, \Omega_3^{\min}, \Omega_4^{\min}) & (\Omega_1^{\max}, \Omega_2^{\min}, \Omega_3^{\max}, \Omega_4^{\max}) \\ (\Omega_1^{\min}, \Omega_2^{\min}, \Omega_3^{\max}, \Omega_4^{\min}) & (\Omega_1^{\max}, \Omega_2^{\max}, \Omega_3^{\min}, \Omega_4^{\max}) \\ (\Omega_1^{\min}, \Omega_2^{\min}, \Omega_3^{\min}, \Omega_4^{\max}) & (\Omega_1^{\max}, \Omega_2^{\max}, \Omega_3^{\max}, \Omega_4^{\min}) \end{array} \quad (4.38)$$

In order to make sure that the solutions of the above-mentioned equations and constraints produce meaningful compounds with reasonable accuracy in properties, a few more rules must be satisfied.

Rule 10: The decision on the groups to be part of a ring or aromatic ring compound should be made ahead of design because the property

contributions of the same group is different in aromatic, cyclic and acyclic compounds. For instance, the property contributions of CH, CH_(ring) and aCH are all different.

Rule 11: The minimum number of molecular fragments forming a ring must be three and for the design of aromatic compounds, there must be exactly six or multiples of six aromatic carbon atoms. If the possible number of aromatic carbon atoms estimated with the first order groups is more than ten, options corresponding to fused ring compounds must also be included, because in that case, the number of aromatic carbon atoms will not be multiples of six.

Rule 12: The number of each group should be a non-negative number.

Rule 13: The Free Bond Number (*FBN*) is the number of free bonds in each molecular string and is represented mathematically by eq. (4.39):

$$FBN = \sum_{g=1}^{N_g} n_g FBN_g - 2 \left(\sum_{g=1}^{N_g} n_g - 1 \right) - 2N_r \quad (4.39)$$

where, N_r is the number of rings (including aromatic groups) in the final molecule and FBN_g is the number of free bonds in each specific group. The molecule's *FBN* should be zero to ensure a complete molecular structure with no charge/free bonds in the final molecule.

The rules can be written mathematically as follows:

$$n_g \geq 0 \quad \sum n_{gr} \geq 3 \text{ or } 0 \quad FBN = 0 \quad \sum n_{ac} = 0,6,10,12,\dots \quad (4.40)$$

where n_{gr} represents the groups forming the ring and n_{ac} is the number of aromatic carbon atoms. The exact numbers for n_{ac} can be written only after identifying the number of first order groups. Constraints must be imposed considering any fused ring compounds along with poly-ring compounds. For instance, if the maximum numbers of aromatic carbon atoms are 16, then the value of $\sum n_{ac}$ can be 6, 10, 12, 13, 14, and 16. The values other than multiples of six correspond to possible fused ring compounds.

4.5.3. Algebraic Molecular Design Algorithm

The procedures used to solve a molecular design problem using the developed algebraic approach are summarized in table 4.2.

Even though the algebraic approach does not have the visual appeal, it is a convenient design tool because of its ability to provide all possible solutions. The algorithm developed can be easily programmed and in the case studies presented in chapter 5, a Visual Basic program was used to solve the equations. Combined with the algebraic treatment of a process design problem (Qin *et al.*, 2004), this approach can be used for process systems with any number of properties to identify the potential molecules with only a medium level of complexity.

Table 4.2: Algebraic Approach Algorithm

<i>Step</i>	<i>Description</i>	<i>Equations</i>
1	Transform the required property range into sets of Maxima and Minima of the normalized molecular property operators using the corresponding functions in GCM.	(4.32)
2	Select the first order groups to form the candidate molecules based on the nature of the final product.	-
3	Select the groups that form aromatic and aliphatic rings (if any).	-
4	Using the contributions of each molecular fragment, develop inequality expressions for each property. At this stage, use inequality expressions from eqs. (4.27) and (4.28). Follow Rule 7 to generate second and third order groups based on the first order groups estimated at this stage.	(4.27)-(4.28)
5	Determine <i>AUP</i> range of the sink	(4.20)
6	Evaluate the FBN_g of all molecular fragments and develop the structural constraints.	(4.40)
7	It is required to solve for open chain and cyclic molecules separately. For open chain compounds, use only those groups corresponding to open chain compounds. Then use the equations obtained through steps 4-6 (except the expression for the number of groups forming a ring) to evaluate the maximum	(4.27)-(4.36)

	possible values of all first order groups. Set the minimum value of all groups as zero. Then, with the obtained values, maximize and minimize the <i>AUP</i> range to get a tighter bound on the search space	
8	Generate all possible combinations of the range of groups obtained from step 7. Then, generate the molecular property operators for each combination. Simultaneously evaluate the <i>FBN</i> and <i>AUP</i> values of all possible combinations. As satisfying <i>AUP</i> is a necessary but not sufficient condition, this step will reduce the search space significantly. Back calculate the properties of those compounds whose <i>AUP</i> is within the range and <i>FBN</i> is zero for verification.	-
9	Select those combinations, which satisfy all constraints. Most of the compounds that satisfy the <i>AUP</i> and <i>FBN</i> constraints will satisfy the property constraints as well.	-
10	Repeat the same procedure for aliphatic and aromatic ring compounds separately by including the molecular fragments forming the rings and all equations obtained through steps 4-6.	-

4.6. Introduction of GC+ Models into the Cluster Space

There are occasions when the GC values of one or more molecular groups to be considered are not available in the existing group contribution data sets. It is also possible that the property contributions corresponding to one property of interest are not available

even if the group is defined for GC approaches. To form property clusters for the groups whose contribution is not defined by GCM, it is possible to employ a property operator defined through the CI method. However, it is to be noted that, the value of bond indices and thus the property contributions will depend on the valence delta value of the atom to which this group is being connected. For instance, if the property of interest is heat of vaporization and one of the potential molecular groups is CHF₂, it is not possible to estimate the property contribution of CHF₂ by the CI method before deciding on the groups forming bonds with CHF₂. The best solution for this problem is to define separate property operators for these kinds of groups, each corresponding to different types of carbon atom that can potentially form bonds with it since the valence delta depends only on the number of hydrogen atoms bonded with that carbon atom. Now, the property operator for the groups estimated through the CI method can be defined as:

$$\psi_{jk}(P_j) = \sum_i (a_{m,i} A_{m,i}) + b(\chi^0)_m + 2c(\chi^1)_{mk} \quad (4.41)$$

$$\Omega_{ijf} = \sum_{g=1}^{N_g} n_g \Omega_{jg1} + \sum_{m=1}^{N_m} n_m \Omega_{CI} \quad (4.42)$$

Here, k is the number of valence delta values of atoms that can form bonds with this group and Ω_{CI} is the normalized property operator corresponding to the molecular operator formed from the CI group. In this approach, there will be more than one molecular property operator corresponding to each group. Once the molecular property

operators are formed, the rest of the operators can be defined exactly the same way as they are defined for GC properties.

4.6.1. GC⁺ Algorithm for Visual Solution of a Molecular Design Problem

The new property models can be applied to both visual and algebraic approaches with some modifications to the existing algorithms. The stepwise procedure for including the molecular property operators based on CI in the visual solution method is explained below:

1. The property targets identified through the process design must be converted to property clusters using eqs. (4.1)–(4.5).
2. Estimate the lower and upper bounds of *AUP*.
3. The feasible property region can be represented on a ternary cluster diagram according to the algorithm provided in sections 4.3 and 4.4. Six unique points can represent the boundaries of this feasibility region. The normalized property values corresponding to these six points are given below

$$\begin{aligned} & \left(\Omega_1^{\min}, \Omega_2^{\min}, \Omega_3^{\max} \right) \left(\Omega_1^{\min}, \Omega_2^{\max}, \Omega_3^{\max} \right) \left(\Omega_1^{\min}, \Omega_2^{\max}, \Omega_3^{\min} \right) \\ & \left(\Omega_1^{\max}, \Omega_2^{\max}, \Omega_3^{\min} \right) \left(\Omega_1^{\max}, \Omega_2^{\min}, \Omega_3^{\min} \right) \left(\Omega_1^{\max}, \Omega_2^{\min}, \Omega_3^{\max} \right) \end{aligned} \quad (4.43)$$

Plot these points on the ternary diagram and the region inside the hexagon formed by these six points will be the feasibility region for the sink.

4. Generate the first order molecular property operators.

5. For non-GC groups, identify the possible types of atoms (based on hydrogen suppressed molecular graphs) that can form bonds with it and estimate valence delta and bond index values. Calculate the zero order and first order CI values.
6. Generate the molecular property operators based on CI. Separate operators have to be calculated for different types of bonds. As the molecular operators formed from these groups differ only in the bond index, their values will be in a very close range. Therefore, it is possible to form a locus of points on the ternary diagram, which can give insights about the possible combinations with other groups for a valid solution.
7. Obtain the normalized molecular property operators.
8. Calculate *AUP* values of all groups.
9. Calculate the molecular property cluster values for all the groups.
10. Plot all the molecular groups on the ternary cluster diagram.
11. Mix different molecular groups according to the procedure presented in section 4.4. When mixing a CI group with a GC group, the number of hydrogen atoms bonded with the GC group will define the corresponding group in the CI locus. The CI group corresponding to the same number of hydrogen atoms in the GC group must be chosen for mixing. In the example shown in figure 4.7, the designer wants to mix a CH₂CO group with a CHF₂ group. The heat of vaporization value of the CHF₂ group is not available in literature. So, according to the GC⁺ method, the cluster values for CHF₂ groups which form bonds with carbon atoms with different numbers of hydrogen are estimated and plotted on the ternary diagram. As the carbon atom in CH₂CO has two hydrogen atoms bonded

to it, the cluster corresponding to $(\text{CHF}_2)\text{-C}$ (with two H atoms) is to be selected from the locus of CHF_2 groups.

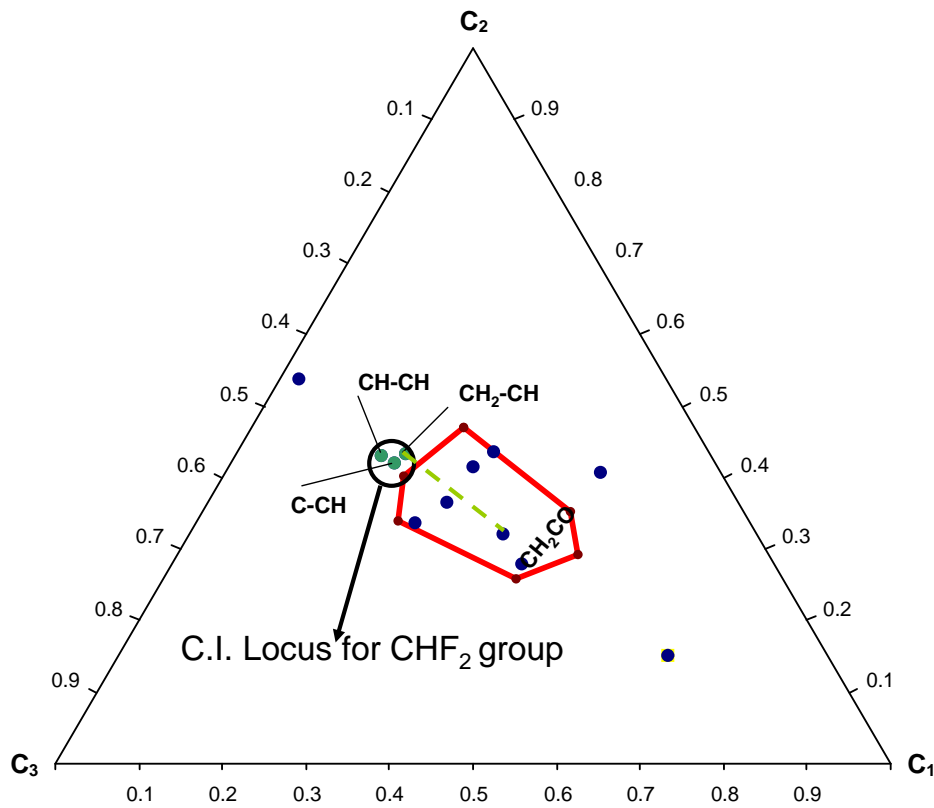


Figure 4.8: Mixing of CI Group with GC Group

12. The possible solutions are those formulations with zero FBN and with an AUP value inside the range set by process targets at the same cluster location defined by the feasibility region.

4.6.2. GC⁺ Algorithm for Algebraic Solution of a Molecular Design Problem

By following the algebraic method, it is possible to include higher order molecular property operators in the design space. To include the CI operators and third order groups into the design space, the procedure given in table 4.3 can be followed.

Table 4.3: GC⁺ Model Algorithm

<i>Step</i>	<i>Description</i>	<i>Equations</i>
1	Select the first order groups to form the candidate molecules	
2	If GC data is not available for any of the groups/properties, estimate it using the CI method for different bond combinations as explained in section 4.3.1. Select the smallest value of the potential group contribution for calculating the property operator to ensure that none of the potential group combinations is missing. Here, the number of potential solutions is overestimated.	
3	Select the groups which form parts of aromatic and aliphatic rings	
4	Develop inequality expressions for each property using the property contributions from each group (GC or CI). Identify the FBN_g of each group and generate the structural constraints. From the inequality expressions for the property constraints and structural constraints, obtain the maximum number of possible groups. With the obtained values, evaluate the maximum and	(4.26)-(4.42)

	minimum values of <i>AUP</i> as well. Identify the possible second and third order groups from these first order groups.	
5	Generate all possible combinations of first order groups. Then, generate molecular property operators for each combination	(4.27)-(4.28)
6	Simultaneously estimate the <i>FBN</i> and <i>AUP</i> values of all combinations. Back calculate the properties of those molecules, which satisfy <i>AUP</i> and <i>FBN</i> constraints to ensure that all the requirements are satisfied.	(4.27)-(4.36) & (4.42)

4.7. Molecular Signatures in Reverse Problem Formulations

The molecular signatures have been shown to produce meaningful QSPR/QSARs and the performance of these descriptors is comparable to many of the existing TIs (Visco *et al.*, 2002; Faulon *et al.*, 2003b). The reason for this correlation is due to the fact that many TIs can be derived from the signature of the molecule. Faulon *et al.* (2003b) have provided the relationships between many TIs and molecular signatures. The general relationship between a TI and its signature has been expressed as a dot product between the vector of the occurrence number of the atomic signature of height *h* and the vector of TI values computed for each root of those atomic signatures:

$$TI(G) = k^h \alpha_G \cdot TI(\text{root}(\sum)) \quad (4.44)$$

Here, k is a constant, ${}^h\alpha_G$ is the vector of the occurrence number of the atomic signature of height h and TI (root (${}^h\Sigma$)) is the vector of TI values calculated for each root of atomic signature. Additional important relations are given in section 4.7.1 & 4.7.2 below.

4.7.1. First Order Connectivity Index

The first order connectivity index is defined using signatures in eq. (4.45):

$${}^1\chi = \frac{1}{2} \sum_{i=1}^{K_G} {}^2\alpha_i \sum_{u \in {}^1V_2({}^hX_i)} [\deg(u) \deg({}^1v_\sigma(u))]^{-1/2} = \frac{1}{2} \sum_{i=1}^{K_G} {}^2\alpha_i \cdot {}^2\Sigma \quad (4.45)$$

Here, ${}^1\chi$ is the first order connectivity index, $\deg(u)$ is the degree of the offsprings and $\deg({}^1v_\sigma(u))$ is the degree of the root atom. The required signature height is two. For instance, the ${}^1\chi$ value of the molecule $\text{CH}_3\text{CH}(\text{CH}_3)\text{CH}_2\text{C}(\text{CH}_3)_3$ is 3.4165. The calculation of connectivity index is as shown below:

$${}^1\chi = (1 \times 3)^{-0.5} + (1 \times 3)^{-0.5} + (3 \times 2)^{-0.5} + (2 \times 4)^{-0.5} + (1 \times 4)^{-0.5} + (1 \times 4)^{-0.5} + (1 \times 4)^{-0.5} = 3.4165 \quad (4.46)$$

Equation (4.45) can be used on the same molecule as follows:

Table 4.4: CI Calculation using Signatures

<i>Signatures of Height 2</i>	<i>Number of Occurrences (2a_i)</i>	${}^2\Sigma$	${}^2\alpha_i \cdot {}^2\Sigma$
C(C(CC))	2	$3^{-0.5}$	1.1547

C(CCC(C))	1	$3^{-0.5}+3^{-0.5}+6^{-0.5}$	1.5629
C(C(CC)C(CCC))	1	$6^{-0.5}+8^{-0.5}$	0.7618
C(CCCC(C))	1	$4^{-0.5}+4^{-0.5}+4^{-0.5}+8^{-0.5}$	1.8536
C(C(CCC))	3	$4^{-0.5}$	1.5

$${}^1\chi = \frac{1}{2}(1.1547 + 1.5629 + 0.7618 + 1.8536 + 1.5) = 3.4165$$

4.7.2. Kier-Hall Shape Index of Order 1

The Kier-Hall shape index, denoted by 1K , can be written in terms of signatures of height 1:

$${}^1K = \frac{(\sum {}^0\alpha_i)([\sum {}^0\alpha_i]-1)^2}{\left[\frac{1}{2}\sum {}^h\alpha_i |{}^1V({}^hX_i)\right]^2} \quad (4.47)$$

Here, the denominator provides the number of paths. Higher orders of shape indices can also be calculated in a similar manner.

The connectivity index and shape index are two of the most widely used TIs in QSAR/QSPR (Trinajstić, 1992). Many other TIs can also be derived from signatures though some of those expressions may be valid only for alkanes. The reason for this kind of relationship becomes clear by analyzing the structure of signature descriptors. Even though the signatures can be considered as the independent building blocks of the molecule, they also depend upon the rest of the signatures of the molecule because each

signature is written in terms of their neighbors. Therefore, the number of building blocks required to represent a fixed number of UNIFAC groups will be more than the actual number of groups. However, the interdependency of such building blocks provides a powerful tool to linearize a variety of highly non-linear topological indices.

Consider the molecular graph of molecule G , of n atoms and diameter D (the maximum chain length possible in the structure). It has been proven that the signature height required to obtain the adjacency matrix of the molecular graph is $h=D+1$. This leads to the conclusion that $D+1$ is the maximum signature height needed to compute any topological index (Faulon *et al.*, 2003b). Faulon *et al.* (2003b) also provided a list of topological indices and the required signature height to represent those topological indices. Some important results from that list is shown in table 4.5:

Table 4.5: Signature Equivalent of Topological Indices

<i>Height</i>	<i>Topological Indices</i>
0	Number of atoms, molecular formula, molecular weight
1	Number of bonds, cyclomatic number, molecular walk count of length 1, shape indices of length 1, connectivity indices of length 0
2	Shape indices of height 2, Platt number, shell index of height 1, connectivity indices of height 1
L	Shape indices of length L, connectivity indices of height L-1

One of the attractive advantages of signature descriptors over many other TIs is its direct application in reverse problem formulations. The reconstruction of the actual

molecule from the solution of a reverse problem is a challenging issue with most TIs. Nevertheless, algorithms are available to enumerate the actual molecular structure from the signatures and the degeneracy of signatures is less than other TIs (Faulon *et al.*, 2003a). Therefore, signature descriptors have the potential to form useful tools in molecular design.

4.7.3. Reverse Problem Formulation using Molecular Signatures

In the previous sections, different algorithms were developed for the simultaneous consideration of both process and molecular design problems. The property models used for the molecular design step in those algorithms were based on group contribution methods. Property models of group contribution methods have attractive qualities making them useful for reverse problem formulations. The property functions can be expressed as linear expressions of the constituent groups in a molecule and the property contributions of the groups are independent of the final molecule. However, the applicability of group contribution methods is limited to a few properties and the property contributions of all the molecular groups the designer wants to consider while designing a molecule may not be available in literature. The structural information provided by group contribution models is limited as well. Therefore, there is a need for a more general algorithm that can be used to predict molecular structures for a wide range of properties and can provide more structural information than normal group contribution method based algorithms do. Since many of the existing QSAR/QSPR expressions can be re-written in terms of molecular signatures, an algorithm based on signatures as the building blocks can meet these targets. However, because of the interdependency of signatures, connectivity rules

must be developed to ensure minimum degeneracy once the solution is generated. Since different topological indices can be represented in terms of signatures, the problems can be solved based on one single descriptor. In addition, there are algorithms available to obtain the molecular structures once the solution is available as signatures (Faulon *et al.*, 2003a).

General problem statement

Identify the best molecules/substituents with the best dominant property, which also satisfy a set of property constraints. The set of property constraints on each property can be represented as follows:

$$P_{ij}^{lower} \leq P_{ij} \leq P_{ij}^{upper} \quad j= 1,2,\dots,N_j ; i=1,2,\dots \quad (4.26)$$

where i is the index of molecules and j is the index of properties.

4.7.4. Signature based Algorithm for Molecular Design

The signature-based algorithm has been developed based on the analogous formulation of property operators and molecular signature descriptors. While comparing the equations used for defining the property operators and molecular signatures, it can be observed that both are defined as linear combinations of the constituent elements. The non-linearity will still appear in the generation of property operators of the individual components. However, since the property operators corresponding to the building blocks can be calculated before solving for the unknown components, this non-linearity will not

contribute to the complexity of the solution procedure. Therefore, it is possible to define a molecular property operator from signatures to track properties based on the contributions of atomic signatures. It can be seen from eq. (4.26) that, each property constraint can be expressed as two inequality expressions: one for the lower bound and one for the upper bound. The first step in solving such a problem is to identify the QSAR/QSPR expressions corresponding to all the target properties/activities. However, different topological indices will be translated in terms of different signature heights. In the next step, identify the molecular signature heights corresponding to the TIs used in QSAR/QSPR. It can be seen in section 4.7.6 that by using signature descriptors, TIs of different heights can be used simultaneously to solve the molecular design problem.

The general form of a QSAR/QSPR can be represented using eq. (4.48):

$$\theta = f(TI) \quad (4.48)$$

where, θ is the property function corresponding to property P . Equation (4.44) can be represented in terms of the number of appearances of signatures using eq. (4.49):

$$TI = \sum_{i=1}^N {}^h \alpha_i L_i \quad (4.49)$$

where

$$L_i = TI \left(\text{root} \left({}^h \sum \right) \right) \quad (4.50)$$

Now, the molecular property operators corresponding to each property can be estimated using eq. (4.51):

$$\psi(P) = \sum_{i=1}^N x_i L_i \quad (4.51)$$

This facilitates the formulation of an optimization problem. The dominant property, which is expressed in terms of the number of atomic signatures, can be maximized or minimized subject to the property constraints. Equation (4.26) can be re-written in terms of normalized property operator as:

$$\begin{aligned} & \text{Max/Min } \Omega_j \\ & \Omega_j^{\min} \leq \Omega_{ij} \leq \Omega_j^{\max} \end{aligned} \quad (4.52)$$

Here, Ω_j is the property operator corresponding to the dominant property and Ω_{ij} is the normalized property operator of molecule i .

The combination of signatures that give the best value for the dominant property should also obey a few rules for the formation of a complete structure. These rules will ensure that the signatures selected based on the property constraints will connect to form a connected graph without any free bonds.

Rule 14: The total number of available degrees (valencies) and the vertices

(atoms) of the graph (molecule) should be selected such that the

molecule must be complete without any free bonds in the structure.

Rule 15: The number of bonds in each signature should match with the bonds in the other signatures.

The best signature combination must be formed in such a way that the above rules are obeyed. For rule 14, a basic rule in graph theory known as the *handshaking lemma* is used. This rule states that, the total sum of valencies of all vertices in a graph will be equal to twice the number of edges (Trinajstic, 1992):

$$\sum_{i=1}^N D(i) = 2M \quad (4.53)$$

Here, D is the number of degrees and M is the number of edges.

Since the available information in reverse problems will be based on the numbers of various candidate signatures, it is necessary to have the equations available in terms of the number of signatures rather than the edges. From graph theory, for a graph with R circuits, the number of vertices (V) can be calculated from the number of edges for a simple graph (graphs without multiple edges):

$$V = M + I - R \quad (4.54)$$

$$\sum_{i=1}^N D(i) = 2(V - 1 + R) \quad (4.55)$$

Consider the design of molecules without multiple edges. Assume the maximum valency of the hydrogen-suppressed atoms involved is limited to four, which is the case with most of the atoms considered in this dissertation. Therefore, for a collection of molecular signatures to form a complete molecule, eq. (4.56) must be obeyed:

$$\sum_{i=1}^{n_1} x_i + 2 \sum_{i=1}^{n_2} x_i + 3 \sum_{i=1}^{n_3} x_i + 4 \sum_{i=1}^{n_4} x_i = 2 \left[\left(\sum_{i=1}^N x_i \right) - 1 + R \right] \quad (4.56)$$

Here, n_1, n_2, n_3, n_4 are the numbers of signatures with valency one, two, three and four, respectively. N is the total number of signatures in the molecule.

Equations (4.54)-(4.56) are applicable only in graphs without multiple edges. However, the constituent signatures may have multiple bonds if the design involves the formulation of molecules with multiple bonds. Therefore, eq. (4.56) has been modified to apply for all types of molecules according to eq. (4.57):

$$\sum_{i=1}^{n_1} x_i + 2 \sum_{i=1}^{n_2} x_i + 3 \sum_{i=1}^{n_3} x_i + 4 \sum_{i=1}^{n_4} x_i = 2 \left[\left[\sum_{i=1}^N x_i + \frac{1}{2} \sum_{i=0}^{N_{Di}} x_i + \sum_{i=0}^{N_{Mi}} x_i + \sum_{i=1}^{N_{Ti}} x_i \right] - 1 + R \right] \quad (4.57)$$

where N_{Di} , N_{Mi} and N_{Ti} are the signatures with one double bond, two double bonds and one triple bond respectively in the parent level. For every additional bond on the signature root level, the number of edges increases by 0.5 because each bond will be shared by two signatures. Therefore, each signature with one double bond on the root will

increase the total number of edges by 0.5. Similarly, since there are two additional edges in a triple bond and the signatures with two double bonds on the root, there will be an addition of one additional edge in the molecular graph.

In order to satisfy rule 15, an expression needs to be formulated that ensures that the bonds at each signature height must be consistent with the rest of the signatures so that there will be a *path* connecting all the vertices. In order to differentiate between different types of atoms, graph coloring has been used. The coloring function has to be appropriately selected based on the types of atoms considered and the nature of the final molecule. For instance, consider the design of an alkane molecule. Here, the coloring function is the valency of each carbon atom at all levels. The coloring should start from the root atom to all atoms up to level $h-1$. Figure 4.9 shows the coloring of one atomic signature of height three:

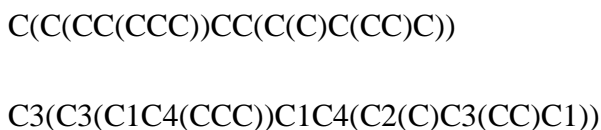


Figure 4.9: Coloring of Atomic Signature

By definition, a molecular graph is an un-directed graph, which is only a function of the vertices and the edges. However, when molecular signatures are used as building blocks to form graphs, they form a *digraph* with respect to the root atom of individual atomic signatures, because not only the individual constituents, but also their directions are significant. In a digraph, the out-degree of one vertex is defined as the number of arcs

formed from the vertex. For a vertex v , the out-degree of v is denoted as $\vec{\rho}(v)$. Similarly, the in-degree of vertex v is the number of arcs joining to the vertex. One of the properties of digraphs known as the *handshaking di-lemma* is useful to ensure the consistency of signatures. According to the handshaking di-lemma, the sum of the in-degrees of all the vertices of the digraph will be equal to the sum of their out-degrees (Wilson, 1986).

$$\vec{\rho}(v) = \vec{\rho}(v) \quad (4.58)$$

When a molecule is formed from signatures, the in-degrees and out-degrees are based on the different types of bonds between atoms. In a complete molecule, since two vertices (atoms) share each edge (bond), the colors of the edge that joins the two vertices must be the same for both the vertices. However, the order of colors will be different for both vertices since the reading of the color has to start from the root atom. For instance, consider the bond formation between the following signatures:

C1(C) and C2(CC)

The edge that joins the above two signatures will have the colors 1 and 2. However, the reading of colors will be 1→2 from the first signature whereas it will be 2→1 for the second signature. The presence of both these edges ensures that there is a linking between the vertices. While writing the atomic signature, only one atom is being described by relating it to its neighboring atoms at different levels of neighborhood. Therefore, each color sequence has to be complemented with another vertex to ensure consistency of the signatures. In other words, every color sequence of edges between any

two heights must be complemented by another signature having one edge with the same colors in the reverse order.

This can be mathematically stated as follows. If $(l_i \rightarrow l_j)_h$ is one coloring sequence $l_i \rightarrow l_j$ at a level h , then, the following equation must be satisfied for the existence of a complete molecule:

$$\sum (l_i \rightarrow l_j)_h = \sum (l_j \rightarrow l_i)_h \quad (4.59)$$

Equation (4.59) has to be obeyed for all color sequences and at each height.

The same rule is to be obeyed even for the color sequence in which $i=j$. That means, if the color sequence in one signature is $l_i \rightarrow l_j$ with $i=j$, then, there must be another signature present in the set of signatures with the same color sequence to complement the previous one. If there are more than one color sequences $l_i \rightarrow l_j$ on one signature with $i=j$, then, all of them must be complemented with the same color sequence in other signatures in the set. For instance, consider the signature of height two, C3(C3(CC)C2(C)C3(CC)). If this signature is present in one molecule, then there must be two more C3→C3 coloring and one C2→C3 coloring in other signatures. This can be mathematically represented as follows:

$$\sum_{i=j} \eta_i x_i = 2K \quad (4.60)$$

where, η is the number of color sequences $l_i \rightarrow l_j$ on one signature with $i=j$ and x is the number of such color sequences and K is an integer.

In order to form a connected tree, it must also be ensured that the total number of signatures where the degree of the vertex at a higher level is more than the degree at a lower level should be less than the total number of vertices with the higher degree. In some signatures, there will be more than one child with a specific color (say m) for a single parent. In such cases, it must be ensured that the number of complementary signatures with the previous parent in the child level must be more than m . For instance, if there is a signature $C2(C3C3)$, then, to form a molecule, there should be at least two distinct signatures with color three in the parent level.

$$\sum x_i n_i \leq \sum x_j \quad (4.61)$$

where, n_i is the number of child vertices with a higher degree than the parent vertex. Here, i and j represent the child and parent colors. Figure 4.9 will explain the above-mentioned principles through an example:



Figure 4.10: Illustration of Connectivity Principles

The height two signatures in this molecule with proper coloring are given below:

C1(C3(CC))

C3(C1C1C2(C))

C2(C3(CC)C3(CC))

C3(C3(CC)C2(C)C1)

C1(C3(CC))

C1(C3(CC))

C3(C3(CC)C1C1)

C1(C3(CC))

C1(C3(CC))

The colors of the edges and their occurrence number are shown in table 4.6:

Table 4.6: Consistency of Signatures

<i>Color Sequence of Edge</i>	<i>Occurrence Number</i>
1→3	5
2→3	2

3→3	2
3→1	5
3→2	2

It can be seen that eqs. (4.59)-(4.61) are satisfied. The total number of degrees for the root carbon atoms is sixteen. The number of vertices is nine. Therefore, eqs. (4.55) and (4.56) are also satisfied. As satisfying these equations provide the necessary and sufficient conditions for the existence of a connected graph, the signatures confirm the existence of a complete molecule.

Now, the dominant property function can be maximized or minimized subject to the constraints in eq. (4.52). The signatures obtained will form the best molecule. In order to form other feasible molecules, integer cuts can be used. Every time a solution has been found, an integer cut can be applied to make sure that the obtained solution will not appear again. This process can be continued until no feasible solution is found, which indicates that all feasible signature combinations that provide molecules satisfying all property constraints have been identified.

The final step is to enumerate the molecular structures corresponding to the identified signatures. An algorithm has been published to generate the molecules once the signatures are available (Faulon *et al.*, 2003a). A simplified algorithm has been developed in this dissertation in section 4.7.7. to generate the molecular structure from the signatures.

4.7.5. Expression of Group Contribution Models with Signatures

Group contribution methods (GCM) have been widely employed to estimate properties of compounds from molecular structures. In 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. Molecular signatures of sufficient height can be used to re-write group contribution expressions. This is because every molecular group can be considered as a tree with $D < h$, where h is the signature height used to represent the molecular groups. Therefore, signatures of that height will be able to describe the atoms, bonds and nearest neighbors.

It should be noted that, if all the property targets were being tracked with group contribution models, the algorithms presented in sections 4.5 and 4.6 would be more efficient. This transformation will be useful only when the property models of some of the target properties are available in GCM and some of the models are available as QSAR/QSPR. If some of the properties of interest are available in GCM, rewriting those models in the form of signatures will allow us to solve the property models based on TIs along with the GC models.

The height and number of the signatures used to write the molecular groups in the GC model depends on the number of atoms used for the molecular design and the nature of final molecule. Different coloring functions can be used to identify the signatures corresponding to the equivalent groups in GC models.

Consider the design of a molecule with amide and alkyl carbon groups. Suppose, the maximum number of amide groups on each alkyl group and attached carbon atom on

each amide group is limited to one. The potential groups available in GCM are (Marrero & Gani, 2001):



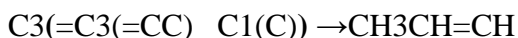
The root atom on every signature can be colored with two numbers: the first color is the number of neighboring C atoms and the second color is the number of neighboring N atoms. Signature of height 2 is required for complete coloring. Now, all signatures with root N can be used to re-write the different amide groups. If the neighboring C atom of an N root has three neighbors, then, the root N atom will be equivalent to a CNH₂ group. Similarly, all N root signatures can be assigned the property contributions of CH₂NH₂ and CHNH₂ groups based on the colors of their nearest neighbors. The signatures with root C atoms will form the alkyl groups. In our example, the signatures with root C atom having N in any of its nearest neighbor should not be considered as a group since that signature has been taken care of in the amide groups. For the rest of the signatures with root C, property contributions can be assigned based on the numbers of neighboring C atoms. Every C root with one neighbor will be equivalent to a CH₃ group since the rest of the valencies are filled by H atoms which are not shown in molecular graphs. Similarly, all C root signatures can be assigned the property contributions of CH₂, CH and C groups based on their colors.

The most important application of signature descriptors while using it to represent group contribution models is its ability to account for the contributions of the higher order molecular groups. As discussed in the previous sections, higher order group effects

are due to the proximity of various first order groups. Therefore, similar to the first order groups, second order groups can also be considered as a tree with $D < h-1$. The procedure to track the second order contributions is as follows:

In the first step, the signatures are generated only based on the first order groups, without considering the second order group contributions. Then, among the generated signatures, identify those signatures that carry the second order group contribution. While assigning property contributions to those signatures, apart from the contribution of the actual molecular group, assign the contribution of the second order group as well. It can be seen that, there is no specific signature for any first order/higher order group. Based on the available groups and the nature of the final molecular structure, the designer can identify the corresponding signature to each molecular group.

The following examples are taken from the solution of the problem presented in section 5.4 of the case study. The first order molecular groups involved are CH₃, CH₂, CH, CH₃CO and CH=CH. After generating the signatures of height three corresponding to the first order molecular groups, the following signatures are identified to carry the second order contributions corresponding to the second order groups shown on the right hand side.



For instance, the first signature has the root color C4, which indicates the carbon with valency four in a hydrogen-suppressed graph. The atoms in the first layer are

oxygen with a double bond, a carbon with valency three and a carbon with valency one. With the available first order groups, this signature distinctly represents a CH₃CO group, with a CH group in its proximity. Similar logic can be seen in the second example as well where the CH=CH group is connected to a CH₃ group.

In general, it has been established that the signature height needed to represent second order group can be two or three. Therefore, it can be concluded that, the maximum signature height required to identify the higher order effect is three.

4.7.6. Property Models with Different Signature Heights

In a molecular design problem with multiple property constraints, it is possible to have different TIs describing different properties of interest. It is also possible to have one QSAR/QSPR containing different TIs. If the heights of the topological indices are different, the signatures corresponding to the largest height have to be enumerated first and the signatures of smaller height have to be represented in terms of the larger signature. This is possible because the total number of any smaller signature can be expressed as the sum of a certain number of higher signatures. For the molecular design algorithm using signatures explained in section 4.7.4, only the number of each signature is significant.

Consider the situation where the QSAR/QSPR has both zero order and first order connectivity indices in the available relationship:

$$Y = a_0 {}^0\chi + a_1 {}^1\chi \quad (4.62)$$

In this case, the corresponding signatures are of heights one and two respectively. Therefore, the signatures will be initially formed corresponding to height two and classified in terms of height one for writing the expression for zero order connectivity index. Here, the property operator for property Y can be written in terms of signatures as follows:

$$f(Y) = \sum_i L_i^1 \alpha_i + \sum_j L_j^2 \alpha_j \quad (4.63)$$

Now, classify all the signatures of height two based on the color of its root. In general, if the height of the largest signature of interest is h and the height of the lower signature is $h-m$, then the classification of signature h is to be done at a level of $h-m$. Now, the total number of each $h-m$ level signatures can be obtained by adding the h level signatures under same color at level $h-m$. This is possible because we are interested only in the number of appearances of each signature in the molecular structure. Since the number of appearances of signatures of lesser height will be represented in terms of the highest signature height, the solution will be obtained in terms of the signatures of the highest height. Therefore, this approach will not increase the degeneracy during the enumeration step. For example, assume the signatures of interest are N1(C), N2(CC) and N3(CCC) which are signatures with height one. Now, signatures of height two with root vertex N can be divided into three sets as follows:

Set 1: Signatures of height two with N vertex and vertex color 1

Set 2: Signatures of height two with N vertex and vertex color 2

Set 3: Signatures of height two with N vertex and vertex color 3

The signatures of height one can be obtained as follows:

$$N(C): {}^1\alpha_{N(C)} = \sum_{\text{set 1}} {}^2\alpha_i$$

$$N(CC): {}^1\alpha_{N(CC)} = \sum_{\text{set 2}} {}^2\alpha_i$$

$$N(CCC): {}^1\alpha_{N(CCC)} = \sum_{\text{set 3}} {}^2\alpha_i$$

Since this methodology solves the problem in terms of the number of appearances of the highest signature in the system, the accuracy will not be sacrificed due to this transformation.

4.7.7. Enumeration of Molecular Structures from Signatures

The generation of molecular graphs from the molecular descriptors is one of the most challenging issues in inverse design. Detailed descriptions of the different approaches have been given in chapter two. However, it is possible to generate the molecular graph from a given set of signatures. In this research project, an algorithm has been developed based on the graph signature enumeration algorithm by Faulon *et al.*, (2003a). This algorithm has been developed to generate the molecular structures from the signature building blocks.

Stepwise procedure

1. Select any signature of height h randomly from the solution set.

2. Consider the signature starting from the first layer of the selected signature (the signature/signatures with height $h-1$). This signature must form the first $n-1$ layers of the signature attached next to the first signature.
3. Generate signatures of height $h-1$ among the rest of the signatures.
4. Select the signature whose $h-1$ height is the same as the signature starting from the first layer of the first signature.
5. If there is more than one signature that satisfy step 4, consider the last layer of the contesting signatures. The signature whose color in the last layer matches with the $(n-1)^{\text{th}}$ layer of the first signature will then be selected for forming the bond. This is possible because, when a bond is formed between two vertices, the same layer will appear in the second signature at the next level.
6. If there is more than one signature that satisfy step 5, it does not matter which signature is selected for forming the bond. This is because, two signatures with the same height h will form isomorphic graphs.
7. After forming the first bond, repeat the same procedure for the other signatures starting at layer one. All matching signatures will form subsequent bonds to the root signature.
8. Repeat the same procedure in the newly formed levels until all the signatures in the solution set have appeared in the signature chain.
9. The signatures will be replaced with the root atoms in each signature.
10. The hydrogen atoms must be added to satisfy the valencies of all the atoms to complete the final molecular structure.

An example of the developed algorithm is presented below. The collection of signatures presented in this example is one of the solutions obtained for the acid gas removal case study in chapter 5. Solution number 3 is selected since it contains both heteroatoms and multiple bonds:

The set of signatures is

- O1(C2(O1(C)C2(CC))) (i)
- O1(C2(O1(C)C2(NC))) (ii)
- C1(N3(C1(N)C2(CN)C2(CN))) (iii)
- C2(O1(C2(OC))C2(C2(OC)C3(=CC))) (iv)
- C2(O1(C2(OC))C2(C2(OC)N3(CCC))) (v)
- C2(N3(C1(N)C2(NC)C2(NC))C2(C2(NC)C3(=CC))) (vi)
- C2(N3(C1(N)C2(NC)C2(NC))C2(C2(NC)O1(C))) (vii)
- N3(C1(N3(CCC))C2(N3(CCC)C2(CC))C2(N3(CCC)C2(OC))) (viii)
- C2(C2(O1(C)C2(CC))C3(=C3(=CC)C2(CC))) (ix)
- C2(C2(N3(CCC)C2(CC))C3(=C3(=CC)C2(CC))) (x)
- C3(=C3(=C3(=CC)C2(CC))C2(C3(=CC)C2(CO))) (xi)
- C3(=C3(=C3(=CC)C2(CC))C2(C3(=CC)C2(NC))) (xii)

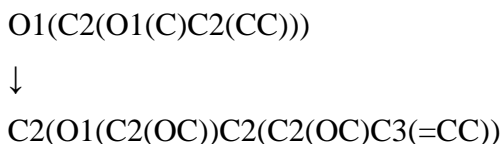
In the first step, select any signature of height 3. In this example, signature (i) has been selected.

O1(C2(O1(C)C2(CC)))

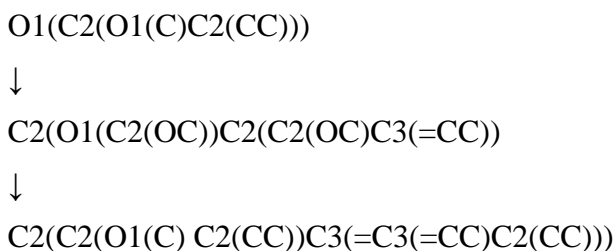
In step two, there is only one signature possible from the first layer, which is

C2(O1(C)C2(CC))

In step 3, all second order signatures of all other signatures have been generated. It can be seen that, for signature (iv), the height two signature is exactly the same as the signature in step two. Since, there are no other signatures of height two, which are the same as the signature in step two, a bond is formed with signature (i) and signature (iv) according to step 4.



The same procedure is repeated on signature (iv) to get the next bond.

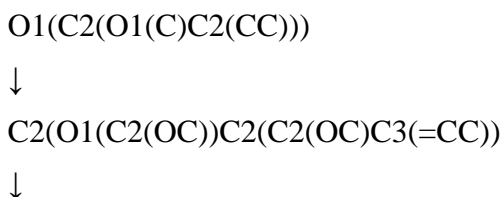


Here, the signature of height two is $\text{C2(O1(C) C2(CC))C3(=C3(=CC)C2(CC))}$.

However, the height two signatures of (xi) and (xii) are the same as this signature:



According to step 5, look for the last layer in such a situation. In signature (xi), an O atom is involved in the final layer whereas an N atom is involved in signature (xii). In the current signature, there is one O atom at its $(n-1)^{\text{th}}$ layer. Therefore, signature (xi) is selected to form the bond:



C2(C2(O1(C) C2(CC))C3(=C3(=CC)C2(CC)))

↓

C3(=C3(=C3(=CC)C2(CC))C2(C3(=CC)C2(CO)))

In any of the following bonds, there are not more than one signature at height three. Therefore, step 6 is not applied in this case. Now, according to steps 7 and 8, the same procedure is repeated until all signatures have appeared in the signature chain. The final signature chain is as follows:

O1(C2(O1(C)C2(CC)))

↓

C2(O1(C2(OC))C2(C2(OC)C3(=CC)))

↓

C2(C2(O1(C) C2(CC))C3(=C3(=CC)C2(CC)))

↓

C3(=C3(=C3(=CC)C2(CC))C2(C3(=CC)C2(CO)))

↓

C3(=C3(=C3(=CC)C2(CC))C2(C3(=CC)C2(NC)))

↓

C2(C2(N3(CCC) C2(CC))C3(=C3(=CC)C2(CC)))

↓

C2(N3(C1(N)C2(NC)C2(NC))C2(C2(NC)C3(=CC)))

↓

N3(C1(N3(CCC))C2(N3(CCC)C2(CC))C2(N3(CCC)C2(OC))) → C1(N3(C1(N)C2(CN)C2(CN)))

↓

C2(N3(C1(N)C2(NC)C2(NC))C2(C2(NC)O1(C)))

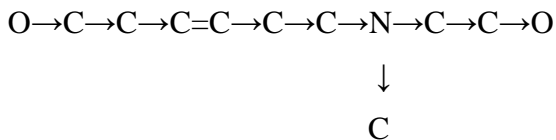
↓

C2(O1(C2(OC))C2(C2(OC)N3(CCC)))

↓

O1(C2(O1(C)C2(NC)))

In step 9, the signatures are replaced with the root atom:



In the last step, the hydrogen atoms are added to get the final molecular structure:



4.7.8. Stepwise Procedure for Solving a Molecular Design Problem

1. Identify the QSPR/QSAR corresponding to the properties of interest.
2. Estimate the height of molecular signatures corresponding to the QSAR.
3. Based on the nature of the target molecule, select the atom types and enumerate the molecular signatures corresponding to the signature height.
4. Re-write the TI in terms of signatures.
5. If different TIs are represented with different heights, express the number of appearances of smaller signatures in terms of number of appearances of larger signatures.
6. Form normalized property operators, which are expressed as linear combinations of atomic signatures.
7. Form the normalized property operators corresponding to GCM if any.
8. The objective function can be defined based on the dominant property.
9. The signature is to be colored at each height up to height $h-1$ with the number of carbon atoms adjacent to it.

10. Form constraints from eqs. (4.56)-(4.61) to ensure the formation of a connected graph and the formation of a complete structure with no free bonds.
11. Solve the objective function corresponding to the constraints and obtain the signatures. Identify the rest of the solutions using integer cuts.
12. Enumerate the molecular structures from signatures according to the procedure in section 4.6.

4.8. General Framework for Integrated Flowsheet and Molecular Design

In chapter two, the development of a group contribution based approach for flowsheet design was introduced. This method can quantify the efficiency of different processing routes from the raw materials to the products. In this section, a general framework is being proposed to integrate flowsheet design with the process and molecular design. The product design should always be conducted simultaneously with the process for which it is being designed because any changes in the process parameters will affect the suitability of the molecules.

The group contribution based approach to estimate the flowsheet property enables one to estimate both how changes in the input molecules and/or the process alternatives affect the processing route. A product identified during the molecular design stage might have superior properties, but may not be suitable because of an unacceptable flowsheet property. For example, a compound with a high value of the energy index for the separation task can make a process economically inefficient because, in most industrial operations, separations account for a large part of the operating cost. However, changing

a product will affect both the process design parameters and the flowsheet design parameters.

In this novel approach, an additional stage has been implemented in the conventional reverse problem formulation framework. The flowsheet property, which is an indication of the efficiency of a process, will be tested after the molecular design stage. However, the flowsheet property will also be a function of the process conditions, as indicated in figure 4.11. Now, in the integrated approach, the flowsheet property will be calculated for each designed molecule. The molecule/molecules within the desirable range of flowsheet properties will be selected for rigorous simulation.

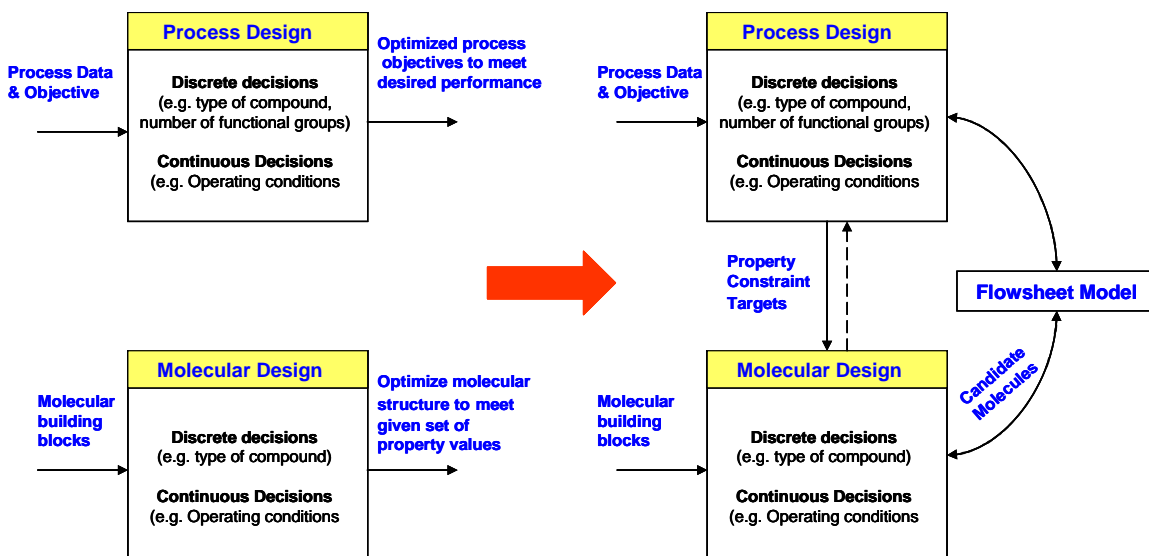


Figure 4.11: Integrated Process-Product-Flowsheet Design Framework

The advantage of this step is, that any molecules that make the process unprofitable can be screened out before any rigorous simulation and experimentation.

The methodology to solve the process and product design problems now can be represented in its entirety as shown in figure 4.12:

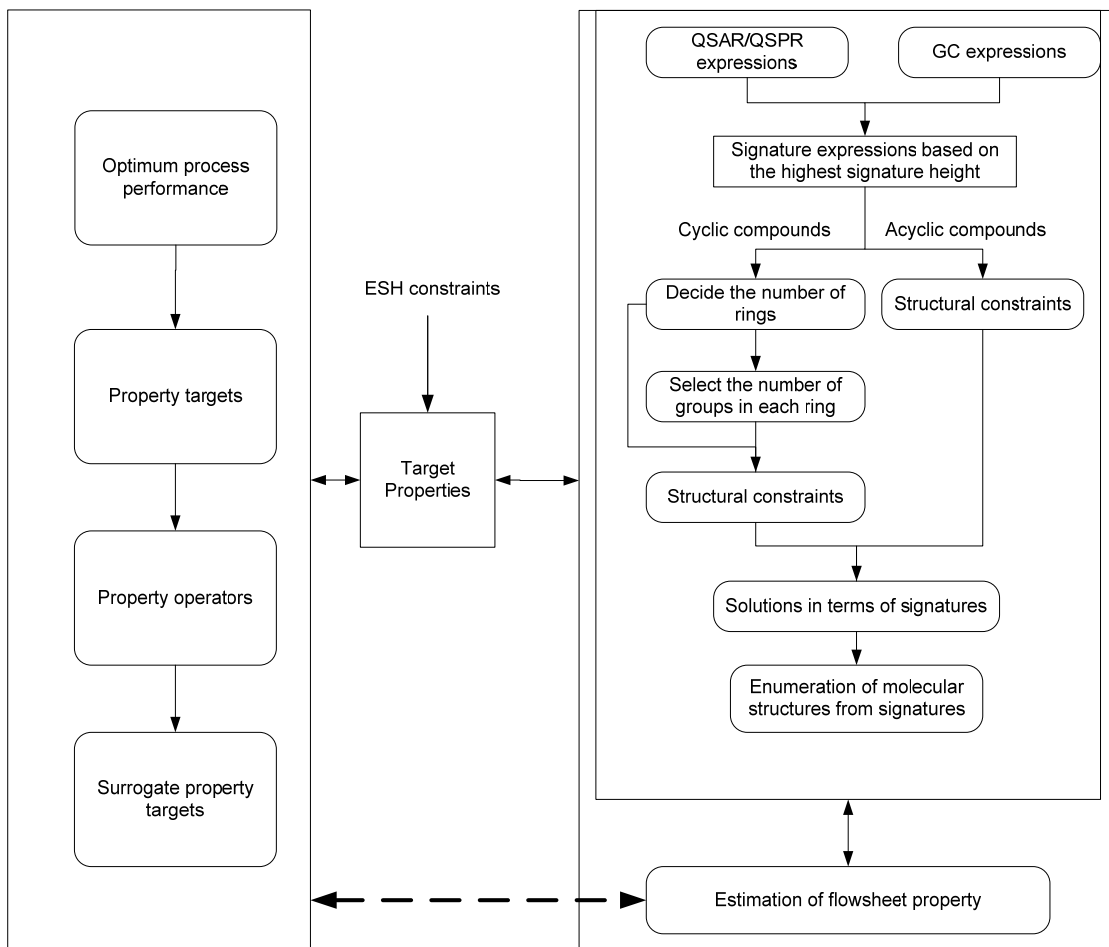


Figure 4.12: Flowchart of the Integrated Process-Product-Flowsheet Design

4.9. Summary

This chapter has covered various algorithms for solving integrated process and product design problems on a property platform. The concept of property clusters has been introduced and the properties of clusters and their visual treatment have been discussed.

The concept has been extended to molecular design by integrating this technique with the property models in group contribution methods. The molecular clustering techniques and their applicability in the simultaneous consideration of process and product design have been discussed. The limitations of the existing methods for molecular design from a property standpoint have been discussed and a new algebraic approach with second and third order levels of molecular groups in Marrero and Gani (2001) models has been introduced. The molecular cluster equations have been redefined to include the higher order effects. A new algorithm has been proposed for the application of these enhanced models in product design. In order to design molecular structures using molecular groups with unknown property contributions, a modified algorithm has been presented that uses a combined connectivity index - group contribution methods approach. The algorithms for both visual and algebraic solution have been presented. The molecular signature descriptors have been included in the reverse problem formulation framework. An algorithm has been developed that can include different QSAR/QSPR expressions based on multiple TIs for molecular design thereby increasing the applicability of the RPF framework. Group contribution methods can be coupled with TI based expressions in the new algorithm on a property platform when the property models are given as both TI based models and group contribution models. In addition, even if the different topological indices are represented with signatures of different heights, the new algorithm utilizes the signature with maximum height to solve the inverse design. The signature-based algorithms have also been integrated with flowsheet design techniques. This general framework will be useful in generating the optimum flowsheet structure for a process on a property platform.

5. Case Studies

In this chapter, four case studies are presented. The first case study is the identification of a blanket wash solvent. The purpose of this study is to highlight the principles involved in the derivation of higher order molecular groups and the systematic procedures followed in the solution of a molecular design problem using the algebraic approach. The second example involves the identification of an alternative metal degreasing solvent, in which the procedure to include connectivity index based groups into the cluster domain is illustrated. The problem is solved both visually and algebraically. In the third case study, a molecular design problem for identifying the most suitable alkyl substituents for a fungicide is solved using molecular signatures. This case study highlights the application of molecular signatures to represent topological indices and group contribution models and the application of connectivity principles. The final case study is a comprehensive integrated process product and flowsheet design problem. An acid gas removal system is studied for the identification of the most effective acid gas removing solvent. The process is evaluated on the basis of performance, environmental constraints and flowsheet properties.

5.1. Design of Blanket Wash Solvent

The application of the developed algebraic approach for product design is illustrated by reworking the design of a blanket wash solvent for a phenolic resin printing

ink. Sinha and Achenie (2001) originally solved this design as a mixed-integer non-linear programming (MINLP) problem and it was later solved visually using the molecular property clusters by Eljack and Eden (2008). In this work, the design has been performed algebraically using molecular property clusters. Group contribution data for the properties considered were taken from Marrero and Gani (2001). The property constraints for the designed solvents are listed in table 5.1.

According to the procedure explained in the section 4.2, the first step in solving a molecular design problem is to convert the property targets into molecular property operators and identify the acceptable ranges for each molecular property operator. The property operators for the given properties and their reference values are listed in table 5.2 and the values of the adjustable parameters are listed in table 5.3. The given property constraints can then be transformed into normalized molecular property operators using the equations in table 5.2 and values in tables 5.2 and 5.3. The calculated values are given in table 5.4.

Seven molecular fragments have been selected for creating the candidate solvents. It should be noted, that the available fragments are the same as those used by Sinha and Achenie (2001). The fragments were selected based on their potential to be a constituent of an industrial solvent. In addition, all these first order groups can form a variety of second order groups, which will be helpful in elucidating the effect of second order GC. The selected groups, their property contributions and number of free bonds are listed in table 5.5.

Table 5.1: Property Constraints for Blanket Wash Solvent

<i>Property (P_j)</i>	<i>Lower Bound</i>	<i>Upper Bound</i>
Standard Heat of Vaporization, H_v	20 kJ/mol	60 kJ/mol
Normal Boiling Temperature, T_b	350 K	400 K
Normal Melting Temperature, T_m	150 K	300 K
Standard Heat of Fusion, H_{fus}	10 kJ/mol	20 kJ/mol

Table 5.2: Property Operators and Reference Values

<i>j</i>	ψ_j	<i>GC Expression</i>	<i>Reference</i>
H_v	$H_v - h_{v0}$	$\sum_{g=1}^{N_g} n_g h_{v1} + \omega \sum_{s=1}^{N_s} n_s h_{v2}$	20
T_b	$\exp\left(\frac{T}{t_{b0}}\right)$	$\sum_{g=1}^{N_g} n_g t_{b1} + \omega \sum_{s=1}^{N_s} n_s t_{b2}$	7
T_m	$\exp\left(\frac{T}{t_{m0}}\right)$	$\sum_{g=1}^{N_g} n_g t_{m1} + \omega \sum_{s=1}^{N_s} n_s t_{m2}$	7
H_{fus}	$H_{fus} - h_{fus0}$	$\sum_{g=1}^{N_g} n_g h_{fus1} + \omega \sum_{s=1}^{N_s} n_s h_{fus2}$	20

Table 5.3: Adjustable Parameters

<i>Adjustable Parameter</i>	<i>Value</i>
h_{v0}	11.733 kJ/mol
t_{b0}	222.543 K
t_{m0}	147.45 K

h_{fus0}	-2.806 kJ/mol
------------	---------------

Table 5.4: Normalized Molecular Property Operator Values

	Ω_{Hv}	Ω_{Tb}	Ω_{Tm}	Ω_{Hfus}
Ω_{min}	0.4134	0.6885	0.3951	0.6403
Ω_{max}	2.4134	0.862	1.0927	1.1403

Table 5.5: Property Data of Selected Molecular Fragments

g	<i>Group</i>	<i>FBN</i>	h_{v1}	t_{b1}	t_{m1}	h_{fus1}
1	CH ₃	1	0.217	0.849	0.695	1.66
2	CH ₂	2	4.91	0.714	0.252	2.639
3	CH	3	7.962	0.293	-0.373	0.134
4	OH	1	24.214	2.567	2.789	4.784
5	CHO	1	12.37	2.539	3.019	11.33
6	CH ₃ CO	1	15.195	3.118	2.959	8.062
7	CH ₂ CO	2	19.392	2.676	2.523	8.826
8	(CH ₂) _{ring}	2	3.341	0.823	0.57	1.069
9	(CH) _{ring}	3	6.416	0.595	0.034	2.511

Now, eq. (4.27) is used to generate the inequality expressions for each property. It should be noted that, only first order groups are considered at this stage. This is because, the expressions generated at this stage are used to obtain the maximum possible number of each groups. The variations in the properties caused by the second order groups will be

considered in the later stages. First equations for the open chain compounds are generated:

$$\begin{aligned}
 0.413 &\leq 0.011g_1 + 0.246g_2 + 0.398g_3 + 1.211g_4 + 0.619g_5 + 0.76g_6 + 0.97g_7 \\
 &\leq 2.413 \\
 0.689 &\leq 0.121g_1 + 0.102g_2 + 0.042g_3 + 0.367g_4 + 0.363g_5 + 0.445g_6 + 0.382g_7 \\
 &\leq 0.862 \\
 0.395 &\leq 0.099g_1 + 0.036g_2 + 0.053g_3 + 0.398g_4 + 0.431g_5 + 0.423g_6 + 0.361g_7 \\
 &\leq 1.093 \\
 0.64 &\leq 0.083g_1 + 0.132g_2 + 0.007g_3 + 0.239g_4 + 0.566g_5 + 0.403g_6 + 0.441g_7 \\
 &\leq 1.14
 \end{aligned} \tag{5.1}$$

The *AUP* range can be estimated from the normalized property operators given in table 5.4. The values are:

$$\begin{aligned}
 AUP_{\min} &= 2.137 \\
 AUP_{\max} &= 5.335
 \end{aligned}$$

From eqs. (4.39) and (4.40), and from the *AUP* range, the structural constraint expressions can be generated as follows:

$$g_1, g_2, \dots, g_7 \geq 0 \tag{5.2}$$

$$\begin{aligned}
 2.137 &\leq 0.3145g_1 + 0.5154g_2 + 0.4933g_3 + 2.2143g_4 + 1.9787g_5 + 2.031g_6 \\
 &+ 2.1537g_7 \leq 5.335
 \end{aligned} \tag{5.3}$$

$$\begin{aligned}
 g_1 + 2g_2 + 3g_3 + g_4 + g_5 + g_6 + 2g_7 - 2(g_1 + g_2 + g_3 + g_4 + g_5 + g_6 + g_7 - 1) \\
 = 0
 \end{aligned} \tag{5.4}$$

Equation (5.1) is used to generate an overview of the potential molecular structure. All the variables in eq. (5.1) are maximized separately subject to the structural constraints in eqs. (5.2)-(5.4). The maximum values are as follows:

$$g_1=4 \quad g_2=6 \quad g_3=3 \quad g_4=1 \quad g_5=2 \quad g_6=1 \quad g_7=1$$

Next, the minimum value for all groups is set to zero. To get a closer bound on the *AUP* values, maximize and minimize the *AUP* subject to the maximum possible groups. The new *AUP* range is

$$AUP_{\min} = 2.875$$
$$AUP_{\max} = 5.005$$

Now, the second order groups that can be formed from the selected molecular fragments are estimated. They, along with their property contribution are listed in table 5.6. It can be seen that second order group 4 is overlapped by group 8, second order group 5 by group 7, and second order group 6 by group 9. Now, all combinations of the first order groups are generated and using eqs. (4.29)-(4.33), molecular property operators for the second order groups are generated. The *AUP* for each combination is then calculated and the structures whose *FBN* is zero and *AUP* is within the range are selected. The properties of the structures are then back calculated and those within the acceptable range are considered for final selection (depending on other parameters like availability, cost etc). The final possible molecular structures along with their estimated properties are given in table 5.7. It can be seen that the two molecules identified in the work of Eljack and Eden (2008) are generated in this design also (the other molecules identified in that work are using different groups). Nevertheless, since the algebraic approach automatically generates the feasible structures, this method identified a third

possible structure (Pentan-3-one) from the same groups even with an additional constraint.

Table 5.6: Second Order Groups and Their Contributions

<i>S</i>	<i>Group</i>	H_{v2}	T_{b2}	T_{m2}	H_{fus2}
1	(CH ₃) ₂ CH	-0.399	-0.0035	0.1175	0.396
2	CHCH ₃ CHCH ₃	0.532	0.316	0.239	-1.766
3	CH-CHO	-0.55	-0.1286	0.5715	-0.369
4	CH ₃ CH ₂ CO	0.403	-0.0215	-0.0968	0.011
5	CH ₃ CHCO	0.723	-0.0803	-0.6024	1.005
6	CH-OH	-0.206	-0.2825	-0.3489	-0.599
7	CHOHCH ₃ CO	-	-0.2987	0.9886	-
8	CH ₂ OHCH ₃ CO	-	-0.2987	0.9886	-
9	CH ₂ CHOH	-	0.5082	-0.5941	-0.041

Table 5.7: Valid Formulations and Their Properties

<i>Molecule</i>	H_v	T_b	T_m	H_{fus}
CH ₃ CO-CHO (2-Oxopropanal ethane 1:1)	39.3	385	263.6	16.6
CHO-CH ₂ -OH (Hydroxyacetaldehyde)	53.2	392	265.6	15.9
CH ₃ -CHOHCHO (2-Hydroxypropanal)	55.7	392	272.6	14.1
CH ₃ (CH ₂) ₂ COCH ₃ (Pentan-2-one)	36.6	374	206.6	12.3
CH ₃ (CH ₂) ₃ CHO (Pentanal)	39.1	380	220.7	18.1
CH ₃ (CH ₂) ₃ OH (Butan-1-ol)	50.9	381	213.0	11.6

$\text{CH}_3\text{CH}_2(\text{CH}_2\text{CO})\text{CH}_3$ (Pentan-3-one)	36.1	361	214.5	12.4
$\text{CH}_3\text{CH}_2\text{CHCH}_3\text{COCH}_3$ (3Methyl pentan2-one)	40.2	388	190.8	12.9
$\text{CH}_3\text{CH}_2\text{CHCHOCH}_3$ (2-Methyl butanal)	36.5	363	236.6	14.6
$\text{CH}_3(\text{CH}_2)_5\text{CH}_3$ (Heptane)	36.9	382	162.2	12.3
$(\text{CH}_3)_3\text{CH}$ (2-Methyl propane)	39.7	380	212.9	11.1
$(\text{CH}_3)_3(\text{CH})_2\text{CHO}$ (2,3 Dimethyl butanal)	40.1	381	235.2	13.4
$(\text{CH}_3)_3(\text{CH}_2)_4\text{CH}$ (2-Methyl heptane)	401.	399	165.7	11.5
CHO-CHO (Ethanedial)	36.5	361	265.1	19.8

In order to identify the possible ring compounds, the molecular groups in table 5.8 are selected:

Table 5.8: Groups For ring Compounds

<i>g</i>	<i>Group</i>
1	CH_3
2	CH_2
3	CH
4	$(\text{CH}_2)_{\text{ring}}$
5	$(\text{CH})_{\text{ring}}$
6	OH
7	CHO

The property constraints can be written in terms of first order groups as in the case of acyclic molecules. One additional structural constraint is:

$$g_4 + g_5 \geq 3 \quad (5.5)$$

The highest possible values of the first order groups can be found by maximizing the variables. The values are:

$$g_1=4 \quad g_2=4 \quad g_3=2 \quad g_4=5 \quad g_5=4 \quad g_6=1 \quad g_7=1$$

The second order groups from these groups and their property contributions are given in table 5.9.

Table 5.9: Second Order Groups and Their Contributions for Cyclic Structures

<i>S</i>	<i>Group</i>	<i>H_{v2}</i>	<i>T_{b2}</i>	<i>T_{m2}</i>	<i>H_{fus2}</i>
1	(CH ₃) ₂ CH	-0.399	-0.0035	0.1175	0.396
2	CHCH ₃ CHCH ₃	0.532	0.316	0.239	-1.766
3	CH-CHO	-0.55	-0.1286	0.5715	-0.369
4	CH-OH	-0.206	-0.2825	-0.3489	-0.599
5	CH ₂ CHOH	-	0.5082	-0.5941	-0.041
6	(CH) _{cyc} -CH ₃	0.096	-0.121	-0.1326	0.033
7	(CH) _{cyc} -CH ₂	-0.428	-0.0148	-0.4669	-1.137
8	(CH) _{cyc} -CH	0.153	0.1395	-0.3548	2.421
9	(CH) _{cyc} -OH	2.134	-0.3179	1.369	-
10	(CH) _{cyc} -CHO	-	-0.2692	0.5076	-

The same methodology used in the previous step is followed for identifying ring compounds. Using the first order, 2nd order acyclic, and the 2nd order ring GCM estimates of the properties for the solvent design results in the 12 potential candidates in table 5.10.

Table 5.10: Valid Cyclic Compounds and Their Properties

<i>Molecular structure</i>	<i>IUPAC Name</i>	H_v (kJ/mol)	T_b (K)	T_m (K)	H_{fus} (kJ/mol)
	cyclobut-2-ene-1-carbaldehyde	46.692	352.62	211.48	17.12
	cyclopent-2-ene-1-carbaldehyde	50.033	387.34	230.26	18.19
	(2E)-3-cyclopropylprop-2-enal	52.575	356.71	222.59	13.07
	(2E)-3-cyclobutylprop-2-enal	55.916	390.85	240.08	14.14
	3-cyclobut-2-en-1-ylpropanal	51.174	382.44	203.71	18.62
	(3E)-4-cyclopropylbut-3-enal	57.057	386.03	215.4	14.57
	3-cycloprop-2-en-1-ylpropanal	52.32	377.43	171.32	19.06
	2-cyclopropylpropanal	51.08	355.73	218.06	17.51
	2-cyclobutylpropanal	54.42	390	236.06	18.58
	(2-ethylcyclopropyl)acetaldehyde	55.96	378.03	223.45	16.2
	3-cyclopropyl-2-methylpropanal	55.56	385.16	210.63	19.01
	3-methylcyclohexene	44.66	391.41	159.15	10.7

5.2. Metal Degreasing Solvent Design

A case study involving a metal degreasing process has been revisited to illustrate the algorithm for the application of GC⁺ techniques in reverse problem formulations. This case study was initially solved to identify the property targets by Shelley and El-Halwagi (2000) and to identify alternative solvents for the degreaser using a visual approach by Eljack *et al.* (2007b). In this work, the focus will be on identifying candidate molecules using first order groups whose properties cannot be estimated using GCM.

In the metal degreasing process, metal parts are sent to a degreaser that uses an organic solvent. In the initial process, the VOCs evaporating in the degreaser had been eliminated by flaring. It has been proposed to condense the VOCs and reuse them along with fresh solvents. Therefore, the first part of this case study is to estimate the lower and upper bounds of the property constraints and the second part will generate the alternative solvent structures for the degreaser.

The methodology to estimate the property targets was developed by Shelley and El-Halwagi (2000) and further extended by Eden *et al.* (2004). In this work, we are identifying the candidate molecules, which satisfy the property targets identified for a fresh solvent. The properties used to describe the new molecules are heat of vaporization (H_v), vapor pressure (VP) and melting point (T_m) (Eljack *et al.*, 2007b). However, for vapor pressure, no group contribution expression exists. Nevertheless, there is an empirical relationship that can be used to calculate vapor pressure from the boiling point (Sinha *et al.*, 2003):

$$\log VP = 5.58 - 2.7 \left(\frac{T_{bp}}{T} \right)^{1.7} \quad (5.6)$$

Here, T_{bp} is the boiling point of the liquid and T is the temperature at which VP is measured, which is 500K in this example. Therefore, it is now possible to represent all the property constraints in terms of group contribution properties, which are shown in table 5.11.

Table 5.11: Property Constraints for the Degreaser Problem

<i>Property</i>	<i>Lower Bound</i>	<i>Upper Bound</i>
VP (mm Hg)	318	1150
H_v (kJ/M)	50	100
T_m (K)	280	350
T_{bo} (K)	480	540

This case study highlights the design of solvents with at least one SO group in their structure for the illustration of the GC⁺ method because the heat of vaporization property data of the SO group is not available in the open literature. The other groups being considered for designing the molecule are:

CH₃, CH₂, CH, OH, CH₃CO, CH₂CO, aCH, aC

The properties listed in table 5.11 can be estimated using the following group contribution expressions:

$$\Delta H_v = h_{v0} + \sum_{g=1}^{n_g} n_i h_{v1} + \sum_{s=1}^{N_s} n_s h_{v2} + \sum_{t=1}^{N_t} n_t h_{v3} \quad (5.7)$$

$$T_b = t_{b0} \cdot \ln \left[\sum_{g=1}^{n_g} n_i t_{b1} + \sum_{s=1}^{N_s} n_s t_{b2} + \sum_{t=1}^{N_t} n_t t_{b3} \right] \quad (5.8)$$

$$T_m = t_{m0} \cdot \ln \left[\sum_{g=1}^{n_g} n_i t_{m1} + \sum_{s=1}^{N_s} n_s t_{m2} + \sum_{t=1}^{N_t} n_t t_{m3} \right] \quad (5.9)$$

However, the ΔH_v value of the SO group is not available in the property contributions published by Marrero and Gani (2001). To estimate the ΔH_v value, the connectivity index method can be used. Now, it should be remembered that, any property contribution calculated through the CI method not only depends on the atoms in the group, but also on the valence delta of the group to which it is connected. In this example, the SO group will form bonds with two groups as it has two free bonds in the structure. From the possible combinations of the available first order groups, the OH group is not considered for a direct bond with the SO group because it will not form a stable compound. All potential groups that can form bonds with the SO group in this case are carbon atoms, but their valence delta will differ based on the number of hydrogen atoms on those carbon groups. In this example, there are four possible types of carbon atoms, that is aliphatic carbon with 1, 2 or 3 hydrogen atoms and aromatic carbon with zero hydrogen. So, there are ten possible values of ΔH_v for the SO group, which are given in table 5.14.

5.2.1. Visual Solution

If the molecules to be designed are not complicated, the visual solution can give reasonably accurate results. In this approach, only first order groups can be considered in the expression for the property function. The property operators and their reference values are given in table 5.12 and the upper and lower limits for the property operators are shown in table 5.13.

Table 5.12: Property Operators and Reference Values for Degreaser Design

<i>Property</i>	<i>Property Operator</i>	<i>GC⁺ Expression</i>	<i>Reference Value</i>
H_v	$\Delta H_v - h_{v0}$	$\sum_{g=1}^{N_g} n_g h_{v1} + f(Y^*)$	20
T_m	$\exp\left(\frac{T}{t_{m0}}\right)$	$\sum_{g=1}^{N_g} n_g t_{m1}$	7
T_{bo}	$\exp\left(\frac{T}{t_{b0}}\right)$	$\sum_{g=1}^{N_g} n_g t_{b1}$	7

Table 5.13: Normalized Molecular Property Operator Values

	Ω_{H_v}	Ω_{T_b}	Ω_{T_m}
Ω_{min}	1.53	2.16	1.67
Ω_{max}	3.53	2.83	2.83

In the next step, the property targets are converted into their corresponding cluster values and the boundaries of the feasibility region defined in step 3 of section 4.6.1 are

determined. These points are plotted on a ternary diagram and connected to obtain the feasibility region corresponding to the target properties.

Now, the property contributions of the groups of interest are obtained and converted into normalized property operators. However, the property contribution of the SO group for heat of vaporization is not available in literature. To use the CI method to calculate the value of H_v , the values of the bond indices and the zero and first order connectivity indices from the valence delta of S and O atoms are estimated. Here, the SO group has two valence electrons in its structure and two bonds are possible from a SO group. In this case study, there are 4 different types of carbon atoms (with 3, 2 or 1 hydrogen and aromatic C) that can potentially form bonds with the SO group. The estimated values are shown in tables 5.14 and 5.15. Now, eq. (4.41) is used to estimate the contribution of the SO group for heat of vaporization for the different possible bonds which are given in table 5.16. Therefore, there are 12 property operators for the SO group. The operators are normalized and the clusters for all groups are calculated and shown in figure 5.1.

Table 5.14: Atom and Bond Indices

<i>Index</i>	<i>Atom/Bond</i>	<i>Value</i>
δ_v	S	2.667
δ_v	O	6
β^k	SO	16
β^k	S-C ₃ C ₃	2.667
β^k	S-C ₃ C ₂	5.333

β^k	S-C ₃ C ₁	8
β^k	S-C ₃ aC	37.334
β^k	S-C ₂ C ₂	10.667
β^k	S-C ₂ C ₁	16
β^k	S-C ₂ aC	74.668
β^k	S-C ₁ C ₁	24
β^k	S-C ₁ aC	112.001
β^k	S-aCaC	522.673

Table 5.15: First Order Connectivity Indices

<i>Group</i>	χ^1
OS-C ₃ C ₃	0.556
OS-C ₃ C ₂	0.467
OS-C ₃ C ₁	0.428
OS-C ₃ aC	0.332
OS-C ₂ C ₂	0.403
OS-C ₂ C ₁	0.375
OS-C ₂ aC	0.308
OS-C ₁ C ₁	0.352
OS-C ₁ aC	0.297
OS-aCaC	0.271

Table 5.16: CI Property Contributions of CI Groups

<i>Group</i>	H_v
SO-C ₃ C ₃	17.71869
SO-C ₃ C ₂	17.26331

SO-C ₃ C ₁	17.06158
SO-C ₃ aC	16.57947
SO-C ₂ C ₂	16.94132
SO-C ₂ C ₁	16.79867
SO-C ₂ aC	16.45776
SO-C ₁ C ₁	16.68219
SO-C ₁ aC	16.40385
SO-aCaC	16.275

Zero order CI for the SO group: 1.021 (Common for all types of SO groups)

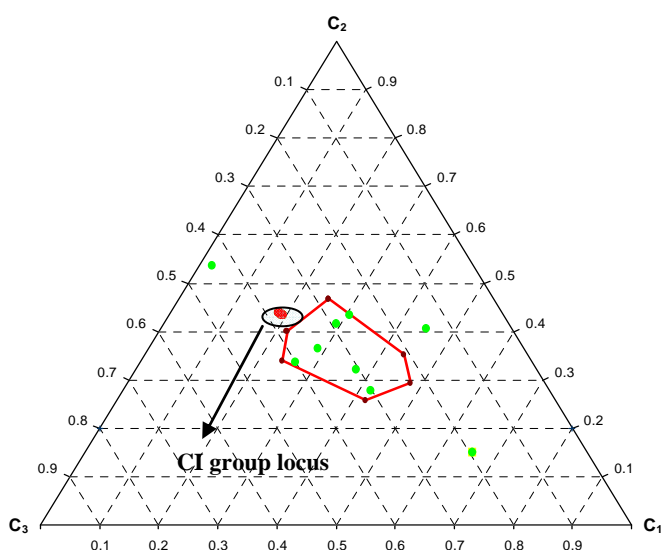


Figure 5.1: Cluster Diagram for Degreaser Design

Note that all cluster locations of the SO groups are close and it is possible to form a locus of SO groups. The reason for such a close range of values is that the major contribution to the property comes from the zero order CI, which depends only on the atoms. Now, the cluster values of different combinations of molecular groups are plotted on the ternary diagram by satisfying the *FBN* constraint defined in eq. (4.39). While

mixing SO groups with other groups make sure that the SO group corresponding to the proper valence delta is used. For instance, if one CH_2CO and one CH_3 are combined with the SO group, the SO group corresponding to carbon atoms with three hydrogen and two hydrogen are to be used. The molecular group formulations, which fall inside the feasibility region and satisfy the *AUP* constraint of the sink are potential solutions. In this problem, some of the identified molecules are shown in the following diagram. The complete solution set is given in the algebraic approach section below.

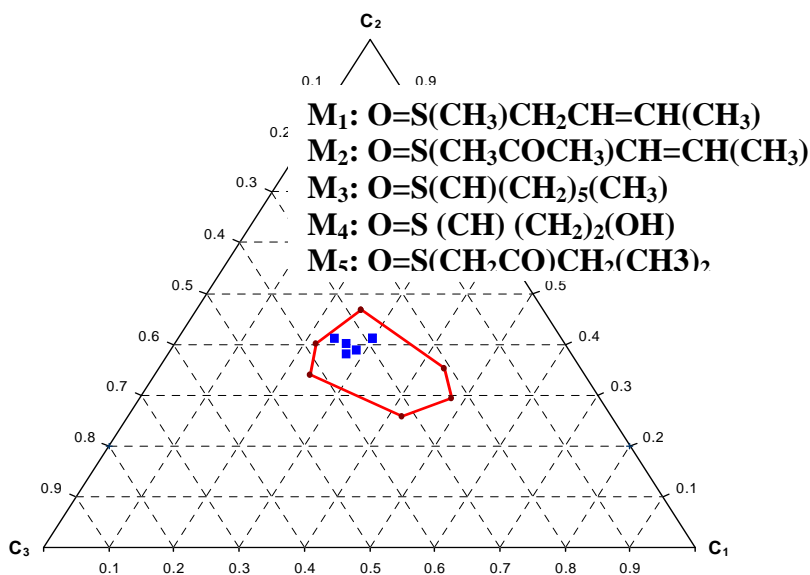


Figure 5.2: Visual Solution of Molecular Design Problem

5.2.2. Algebraic Solution

In this problem, the property contributions of all molecular groups are available except the ΔH_v value of the SO group. In the algebraic approach, the first step is to estimate the maximum possible number of each group. So, for the ΔH_v value of the SO group, the lowest among the estimated values is being considered in the initial stage of

design to make sure that no potential molecule is ignored. From the previous section, this value is 16.68 kJ/mol.

Now, eq. (4.26) is used to generate the inequality expressions corresponding to each property. These equations are used only to estimate the maximum number of each first order group. All these equations are maximized subject to the structural constraints in eqs. (4.36) and (4.39). The maximum values are as follows:

SO:1 aC: 1 aCH: 5 CH₃: 6 CH₂: 8 CH:4 OH:1 CH₃CO:1 CH₂CO:2

The second order and third order groups possible from these first order groups and their property contributions are listed in table 5.17. In table 5.17, the letters *n, m, p* and *k* represent the different values possible for hydrogen.

Table 5.17: Possible higher order groups and their property contributions

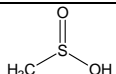
<i>Group</i>	<i>Property Contribution</i>		
	<i>H_v</i>	<i>T_b</i>	<i>T_m</i>
(CH ₃) ₂ CH	-0.399	-0.0035	0.1175
CH(CH ₃)CH(CH ₃)	0.532	0.316	0.239
CH _n =CH _m	1.632	0.1097	0.745
CH _p =CH _k	0.064	0.0369	0.0524
CH ₃ -CH _m =CH _n	-0.06	-0.0537	-0.1077
CH ₂ -CH _m =CH _n	0.004	-0.0093	-0.2485
CH _p -CH _m =CH _n	-0.403	-0.0215	-0.0968
CH ₃ COCH ₂	0.723	-0.0803	-0.6024

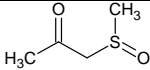
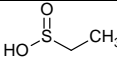
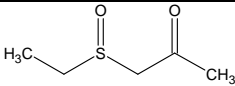
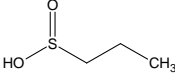
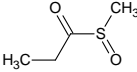
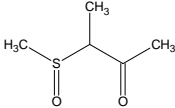
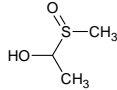
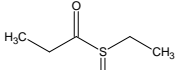
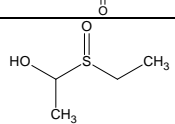
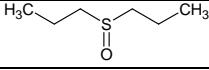
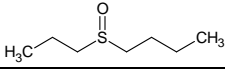
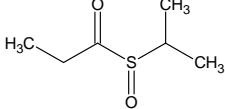
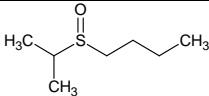
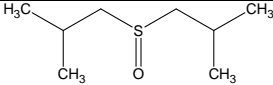
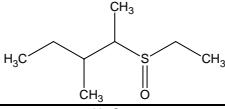
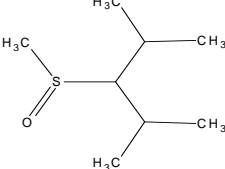
CH ₃ COCH	-	-0.2825	-0.3489
CHOH	-0.206	-	0.9886
CH ₃ COCH _n OH	-	-0.2987	-0.5941
CH _m (OH)CH _n	-	0.5082	-

Now, all possible combinations of first order groups are generated and the property contributions from potential higher order groups are estimated using eqs. (4.27)-(4.35). The molecular property operators are generated subject to the constraints in eqs. (4.39)-(4.40) using eq. (4.36) and the *AUP* values for each combination are calculated. The combinations whose *AUP* values are within the limits are potential solutions. The property values of those combinations are back calculated to confirm they are real solutions. Seventeen molecular structures were identified using this method and their structure along with the predicted properties are given in table 5.18.

It can be seen that all the compounds identified in the visual approach are identified in the algebraic approach as well. The table shows only those structures whose property values fall in the required region based on the GC⁺ method after obtaining the basic structures. Most of the structures obtained by the algebraic approach will satisfy all the property constraints. It should be remembered that the basic purpose of this analysis is to short list the potential candidates and ensure that no possible candidate is missed for further investigation.

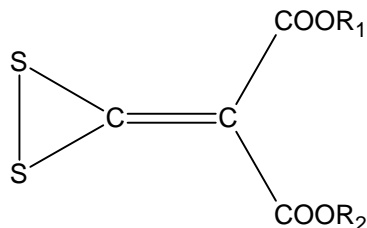
Table 5.18: Final Solution Set for Degreaser Design

<i>Molecule</i>	<i>H_v(kJ/moe)</i>	<i>T_b(K)</i>	<i>T_m(K)</i>
	52.7	505.6	321.5

	50.1	532.4	326.8
	57.6	521.4	325.6
	53.1	539.8	330.8
	62.5	536.1	329.7
	50.1	526.1	326.9
	52.6	539.8	331.8
	60.7	524.4	333.0
	52.6	540	330.9
	65.6	538.9	336.8
	50	530.3	302.2
	53.3	540	306.9
	55.9	539.9	343.1
	51.2	538.8	318.9
	54.2	540	333.9
	50.5	539.5	330.1
	52.5	539.7	342.4

5.3. Design of Alkyl Substituent for the Fungicide DD

The application of molecular signature descriptors in molecular design is illustrated by reworking a case study involving the optimal substituent selection for dialkyldithiolanylidene malonate (DD).



This design problem was solved by Raman and Maranas (1998) by incorporating TIs as structural descriptors. In this work, the molecular signature descriptors are used to solve the molecular design problem. Since the new algorithm can be used to re-write group contribution relationships in terms of molecular signatures as well, an additional property constraint is included to illustrate this capability. Group contribution data for the properties considered were taken the work of Martin and Young (2001).

5.3.1. Problem Statement

DD is a fungicide that has eradicant and protective ability against rice blast disease. The effects of DD is quantified in terms of affinity ($\log(V_E)$), mobility ($\log(\mu)$) and retention ($\log[R/(1-R)]$) in the plant and the correlation of these properties with hydrophobic factor ($\log(P)$) has been published by Uchida (1980). The linear relation between $\log(P)$ and the first order molecular connectivity index ${}^1\chi$ was developed by Murray *et al.* (1975). Therefore, it is possible to correlate affinity, mobility and retention with ${}^1\chi$. The correlations relating these properties to ${}^1\chi$ have been developed by Raman and Maranas (1998) and are shown in eqs. (5.10)-(5.12). An additional property of

interest while designing a fungicide is the toxicity (LC_{50}). Group contribution methods can be used to predict LC_{50} from the molecular structure according to eq. (5.13) (Martin & Young, 2001). The ability to combine different property models in RFP using signatures is another unique contribution of this dissertation.

5.3.2. Solution of Design Problem with Two Types of Property Models

The objective of this case study is to identify the alkyl substituents of DD that give maximum affinity subject to constraints on mobility, retention and toxicity. The property targets have been identified for only the part of the molecule that changes. The upper and lower bounds of the properties are given in table 5.19.

$$\log(V_E) = 0.5751({}^1\chi) - 0.2942 \quad (5.10)$$

$$\log(\mu) = -0.6983({}^1\chi) + 2.0143 \quad (5.11)$$

$$\log\left(\frac{R}{1-R}\right) = 0.787({}^1\chi) - 2 \quad (5.12)$$

$$\log(LC_{50}) = -\sum_{i=1}^N n_i \alpha_i \quad (5.13)$$

Table 5.19: Upper and Lower Bounds for Fungicide Properties

<i>Property (P_j)</i>	<i>Lower Bound</i>	<i>Upper Bound</i>
Mobility	-0.3	0.3

Retention time	-0.3	1.0
LC ₅₀	0.04	-

The first step in solving the molecular design problem is to re-write the property expressions in terms of signatures. The first order connectivity index can be re-written in terms of molecular signatures of height two according to eq. (4.45):

$${}^1\chi = \frac{1}{2} \sum_{i=1}^{K_G} {}^h\alpha_i \sum_{u \in {}^hV_2({}^hX_i)} [\text{deg}(u) \text{deg}({}^{-1}v_\sigma(u))]^{-1/2} \quad (4.45)$$

For simplicity, eq. (4.45) can be re-written as follows:

$${}^1\chi = \sum_{i=1}^{K_G} L_i {}^h\alpha_i \quad (5.14)$$

where

$$L_i = \frac{1}{2} \sum_{u \in {}^hV_2({}^hX_i)} [\text{deg}(u) \text{deg}({}^{-1}v_\sigma(u))]^{-1/2} \quad (5.15)$$

The signature descriptors can be used to re-write group contribution expressions as well. Since the target molecules are alkanes, there will be four first order groups available (C, CH, CH₂ and CH₃). All signatures can be classified into any of these four groups based on the number of suppressed hydrogen atoms on each root. For alkanes, it is

possible to form 65 signatures of height 2 as shown in table 5.20. It can be seen that, three signatures correspond to the CH₃ group, nine signatures correspond to the CH₂ group, nineteen correspond to the CH group and thirty-four correspond to the C group. The reference value is taken as one for all the properties. Therefore, the normalized property operators are as follows:

$$\Omega_{V_E} = \frac{\log(V_E) + 0.2942}{0.5751} \quad (5.16)$$

$$\Omega_{\mu} = \frac{\log(\mu) - 2.0143}{-0.6983} \quad (5.17)$$

$$\Omega_R = \frac{\log\left(\frac{R}{1-R}\right) + 2}{0.787} \quad (5.18)$$

$$\Omega_{LC_{50}} = -\log(LC_{50}) \quad (5.19)$$

The molecular design problem can be written in terms of signatures as follows:

$$\text{Max } \Omega_{V_E} \quad (5.20)$$

$$2.16 \leq \sum_{i=1}^{65} h_i x_i \leq 3.314 \quad (5.21)$$

$$2.445 \leq \sum_{i=1}^{65} h_i x_i \leq 3.812 \quad (5.22)$$

$$3.219 \geq \sum_{i=1}^3 c_1 x_i + \sum_{i=4}^{12} c_2 x_i + \sum_{i=13}^{31} c_3 x_i + \sum_{i=32}^{65} c_4 x_i \quad (5.23)$$

Equations (5.21) and (5.22) can be re-written based on interval arithmetic to eq. (5.24):

$$2.455 \leq \sum_{i=1}^{65} h_i x_i \leq 3.314 \quad (5.24)$$

Table 5.20: Signatures of Height Two for Alkanes

<i>Height Two Signatures</i>	
C1(C2(C))	C4(C4(CCC)C4(CCC)C4(CCC)C1)
C1(C3(CC))	C4(C4(CCC)C4(CCC)C3(CC)C3(CC))
C1(C4(CCC))	C4(C4(CCC)C4(CCC)C3(CC)C2(C))
C2(C2(C)C1)	C4(C4(CCC)C4(CCC)C3(CC)C1)
C2(C3(CC)C1)	C4(C4(CCC)C4(CCC)C2(C)C2(C))
C2(C4(CCC)C1)	C4(C4(CCC)C4(CCC)C2(C)C1)
C2(C2(C)C2(C))	C4(C4(CCC)C4(CCC)C1C1)
C2(C3(CC)C2(C))	C4(C4(CCC)C3(CC)C3(CC)C3(CC))
C2(C4(CCC)C2(C))	C4(C4(CCC)C3(CC)C3(CC)C2(C))
C2(C4(CCC)C4(CCC))	C4(C4(CCC)C3(CC)C3(CC)C1)

C2(C3(CC)C3(CC))	C4(C4(CCC)C3(CC)C2(C)C2(C))
C2(C4(CCC)C3(CC))	C4(C4(CCC)C3(CC)C2(C)C1)
C3(C4(CCC)C4(CCC)C4(CCC))	C4(C4(CCC)C3(CC)C1C1)
C3(C4(CCC)C4(CCC)C3(CC))	C4(C4(CCC)C2(C)C2(C)C2(C))
C3(C4(CCC)C4(CCC)C2(C))	C4(C4(CCC)C2(C)C2(C)C1)
C3(C4(CCC)C4(CCC)C1)	C4(C4(CCC)C2(C)C1C1)
C3(C4(CCC)C3(CC)C3(CC))	C4(C4(CCC)C1C1C1)
C3(C4(CCC)C3(CC)C2(C))	C4(C3(CC)C3(CC)C3(CC)C3(CC))
C3(C4(CCC)C3(CC)C1)	C4(C3(CC)C3(CC)C3(CC)C2(C))
C3(C4(CCC)C2(C)C2(C))	C4(C3(CC)C3(CC)C3(CC)C1)
C3(C4(CCC)C2(C)C1)	C4(C3(CC)C3(CC)C2(C)C2(C))
C3(C4(CCC)C1C1)	C4(C3(CC)C3(CC)C2(C)C1)
C3(C3(CC)C3(CC)C3(CC))	C4(C3(CC)C3(CC)C1C1)
C3(C3(CC)C3(CC)C2(C))	C4(C3(CC)C2(C)C2(C)C2(C))
C3(C3(CC)C3(CC)C1)	C4(C3(CC)C2(C)C2(C)C1)
C3(C3(CC)C2(C)C2(C))	C4(C3(CC)C2(C)C1C1)
C3(C3(CC)C2(C)C1)	C4(C3(CC)C1C1C1)
C3(C3(CC)C1C1)	C4(C2(C)C2(C)C2(C)C2(C))
C3(C2(C)C2(C)C2(C))	C4(C2(C)C2(C)C2(C)C1)
C3(C2(C)C2(C)C1)	C4(C2(C)C2(C)C1C1)
C3(C2(C)C1C1)	C4(C4(CCC)C4(CCC)C4(CCC)C4(CCC))
C4(C2(C)C1C1C1)	C4(C4(CCC)C4(CCC)C4(CCC)C3(CC))
C4(C4(CCC)C4(CCC)C4(CCC)C2(C))	

To make sure that, in the final solution, no free bonds will be present, use eq. (4.56). Here, we are looking for only acyclic substituents because the target molecule is an acyclic alkane. Therefore, the connectivity rules corresponding to the acyclic structure can be used.

$$\sum_{i=1}^3 x_i + 2 \sum_{i=4}^{12} x_i + 3 \sum_{i=13}^{31} x_i + 4 \sum_{i=32}^{65} x_i = 2 \left[\left(\sum_{i=1}^{65} x_i \right) - 1 \right] \quad (5.25)$$

Now, equations need to be formulated to ensure that the identified signatures will connect together completely to form meaningful compounds. To differentiate among different carbon atoms in the structure, vertex coloring is used. Here, coloring of the vertices is performed with the degree of each carbon atom in a hydrogen-suppressed graph. So, according to the procedure explained in section 4.7.4, these signatures must be made consistent with the rest of the signatures in the solution set. In other words, if there is a bond sequence between different colors in a signature, there must be a different signature with the same color sequence in reverse order. The number of any color sequence must be equal to the total number of reverse color sequence. These considerations can be written in algebraic form as follows:

$$\begin{aligned} x_1 &= x_4 + x_5 + x_6 \\ x_2 &= x_{16} + x_{19} + x_{21} + 2x_{22} + x_{25} + x_{27} + 2x_{28} + x_{30} + 2x_{31} \\ x_3 &= x_{35} + x_{38} + x_{40} + 2x_{41} + x_{44} + x_{46} + 2x_{47} + x_{49} + 2x_{50} + 3x_{51} + x_{54} + x_{56} + 2x_{57} \\ &+ x_{59} + 2x_{60} + 3x_{61} + x_{63} + 2x_{64} + 3x_{65} \\ x_5 + x_8 + 2x_{11} + x_{12} &= x_{15} + x_{18} + 2x_{20} + x_{21} + x_{24} + 2x_{26} + x_{27} + 3x_{29} + 2x_{30} + x_{31} \end{aligned}$$

$$\begin{aligned}
x_6 + x_9 + 2x_{10} + x_{12} &= x_{34} + x_{37} + 2x_{39} + x_{40} + x_{43} + 2x_{45} + x_{46} + 3x_{48} + 2x_{49} + x_{50} + x_{53} \\
&+ 2x_{55} + 3x_{58} + 2x_{59} + x_{60} + 4x_{62} + 3x_{63} + 2x_{64} + x_{65} \\
3x_{13} + 2x_{14} + 2x_{15} + 2x_{16} + x_{17} + x_{18} + x_{19} + x_{20} + x_{21} + x_{22} &= x_{33} + 2x_{36} + x_{37} + x_{38} \\
&+ 3x_{42} + 2x_{43} + 2x_{44} + x_{45} + x_{46} + x_{47} + 4x_{52} + 3x_{53} + 3x_{54} + 2x_{55} + 2x_{56} + 2x_{57} + x_{58} \\
&+ x_{59} + x_{60} + x_{61}
\end{aligned} \tag{5.26}$$

Now, eqs. (4.59) and (4.60) are used to ensure the consistency of the signatures with the same color sequence on same edges:

$$\begin{aligned}
x_4 + 2x_7 + x_8 + x_9 &= 2h_1 \\
(x_{14} + x_{18} + x_{19} + x_{26} + x_{27} + x_{28}) + 2(x_{17} + x_{24} + x_{25}) + 3x_{23} &= 2h_2 \\
4x_{32} + 3(x_{33} + x_{34} + x_{35}) + 2(x_{36} + x_{37} + x_{38} + x_{39} + x_{40} + x_{41}) + x_{42} + x_{43} + x_{44} + \\
x_{45} + x_{46} + x_{47} + x_{48} + x_{49} + x_{50} + x_{51} &= 2h_3
\end{aligned} \tag{5.27}$$

In order to ensure that the number of a specific color in the child level would not exceed the total number of the same color when it is in parent level, eq. (4.61) is employed:

$$\begin{aligned}
2x_{11} &< \sum_{3 \rightarrow 2} x_i \\
2x_{10} &< \sum_{4 \rightarrow 2} x_i \\
3x_{13} &< \sum_{4 \rightarrow 3} x_i \\
2(x_{14} + x_{15} + x_{16}) &< \sum_{4 \rightarrow 3} x_i \\
4x_{32} &< \sum_{4 \rightarrow 4} x_i \\
3(x_{33} + x_{34} + x_{35}) &< \sum_{4 \rightarrow 4} x_i \\
2(x_{36} + x_{37} + x_{38} + x_{39} + x_{40} + x_{41}) &< \sum_{4 \rightarrow 4} x_i \\
3x_{23} &< \sum_{3 \rightarrow 3} x_i \\
2(x_{17} + x_{24} + x_{25}) &< \sum_{3 \rightarrow 3} x_i \\
2x_7 &< \sum_{2 \rightarrow 2} x_i
\end{aligned} \tag{5.28}$$

Equations (5.21)-(5.29) can be solved to identify the best signature combination that maximizes V_E . In order to compare the new algorithm with the existing work by Raman and Maranas (1998), the problem is solved initially without considering the toxicity constraint and the solution in terms of signatures is shown in table 5.21. The enumerated molecular structures from these signatures along with the estimated properties are shown in table 5.22 without considering the toxicity constraint. It should be noted that this is exactly the same solution as published by Raman and Maranas (1998). The final solution to this problem subject to all the constraints, including toxicity, is also identified. Now, only three solutions are found to be feasible from the initial list. The first

two solutions in table 5.21/5.22 are no longer valid when all the constraints are considered.

Table 5.21: Solution in Terms of Signatures

<i>Order</i>	<i>Signature</i>	<i>Occurrence</i>
1	C1(C2)	2
	C1(C3)	1
	C2(C2C1)	1
	C2(C3C1)	1
	C2(C3C2)	1
	C3(C2C2C1)	1
2	C1(C2)	1
	C1(C3)	2
	C2(C2C1)	1
	C2(C2C2)	1
	C2(C3C2)	1
	C3(C2C1C1)	1
3	C1(C2)	1
	C1(C3)	3
	C2(C3C1)	1
	C3(C3C2C1)	1
	C3(C3C1C1)	1
4	C1(C3)	4

	C2(C3C3)	1
	C3(C2C1C1)	2
5	C1(C2)	2
	C1(C4)	2
	C2(C4C1)	2
	C4(C2C2C1C1)	1

Table 5.22: Possible Alkyl Substituents

<i>Properties</i>				<i>R1</i>	<i>R2</i>
<i>Affinity</i>	<i>Mobility</i>	<i>Retention</i>	<i>Toxicity</i>		
1.6083	-0.2957	0.6034	0.0353	methyl	3-methyl-butyl
				methyl	2-pentyl
				ethyl	sec-butyl
1.5864	-0.2691	0.5735	0.0353	methyl	iso-pentyl
				ethyl	iso-butyl
				n-propyl	iso-propyl
1.535	-0.2068	0.5032	0.0401	methyl	2-methyl-2-butyl
1.5035	-0.1685	0.4601	0.0408	iso-propyl	iso-propyl
1.5009	-0.1653	0.4565	0.042	methyl	tert-pentyl

5.3.3. Solution of Design Problem with Different Topological Indices

The algorithm developed in this dissertation can be applied to design problems where the property models for different target properties have been predicted using different topological indices. To illustrate its application in property models with more than one topological index, the case study in section 5.3.2 is re-solved with a different constraint. Here, instead of toxicity, toxic limit concentration will be used as a constraint in the molecular design problem. For toxic limit concentration, a QSAR model is available that uses the connectivity index of order two (Koch, 1982):

$$\log(TLC) = 4.204 - 1.385({}^2\chi^v) \quad (5.29)$$

It should be noted that, the molecular design problem has property models with two different topological indices, i.e. connectivity indices of order one and two. Connectivity index of order two can be represented using eq. (5.30):

$${}^2\chi = \frac{1}{2}({}^3\alpha_G)^2 \chi(\text{root}({}^3\Sigma)) \quad (5.30)$$

From eq. (5.30), it is clear that signature descriptors of height three are required to represent the topological index involved in eq. (5.29). Following the procedure presented in section 5.3.2, all significant second order signatures are generated and similar equality

expressions are developed. The lower limit of $\log(TLC)$ is kept as 3. The property operator function and limits are shown in eqs. (5.31) and (5.32), respectively:

$$\Omega_{TLC}=(4.204 - \log(TLC))/1.385 \quad (5.31)$$

$$\Omega < 0.87 \quad (5.32)$$

Now, the first order connectivity indices are represented using signatures of height two and the second order connectivity indices are represented using signatures of height three. In the next step, all signatures of height two are represented with the number of appearances of signatures of height three using eq. (5.33). Here, set 1, 2, 3 and 4 represent the signatures of height three with root vertex color 1, 2, 3 and 4, respectively:

$$\begin{aligned} C(C) : {}^2\alpha_{c(c)} &= \sum_{\text{set 1}}^3 \alpha_i \\ C(CC) : {}^2\alpha_{c(cc)} &= \sum_{\text{set 2}}^3 \alpha_i \\ C(CCC) : {}^2\alpha_{c(ccc)} &= \sum_{\text{set 3}}^3 \alpha_i \\ C(CCCC) : {}^2\alpha_{c(cccc)} &= \sum_{\text{set 4}}^3 3\alpha_i \end{aligned} \quad (5.33)$$

The optimization problem is now solved for the signatures of height three. There is only one solution that satisfies all the constraints:

Table 5.23 New Solution for Alkyl Substituent

<i>Affinity</i>	<i>Mobility</i>	<i>Retention</i>	<i>log(TLC)</i>	<i>R₁</i>	<i>R₂</i>
1.5009	-0.1653	0.4565	1.83	methyl	<i>tert</i> -pentyl

It can be seen that the developed algorithm can be applied to molecular design problems where different property models are involved (in section 5.3.2 where both connectivity index based models and group contribution based models are used) and when different topological indices, that require the transformation of molecular signatures of different heights, are involved (in section 5.3.3 where connectivity indices of order one and two are used, which require the use of molecular signatures of heights two and three).

5.4. Acid Gas Removal

5.4.1. Problem Statement

A gas treatment process uses methyl diethanol amine, MDEA ($\text{OH}(\text{CH}_2)_4\text{N}(\text{CH}_3)\text{OH}$) to remove acid gases from an alkane rich feed. Two recycled process streams (S_1 and S_2), which mainly consists of 2,5 Dimethyl hexane are also in the feed and will be separated from the amine after the acid gas absorption. The property and flowrate data of both MDEA and the recycled streams are summarized in table 5.24. The flowsheet of the process is given in figure 5.3.

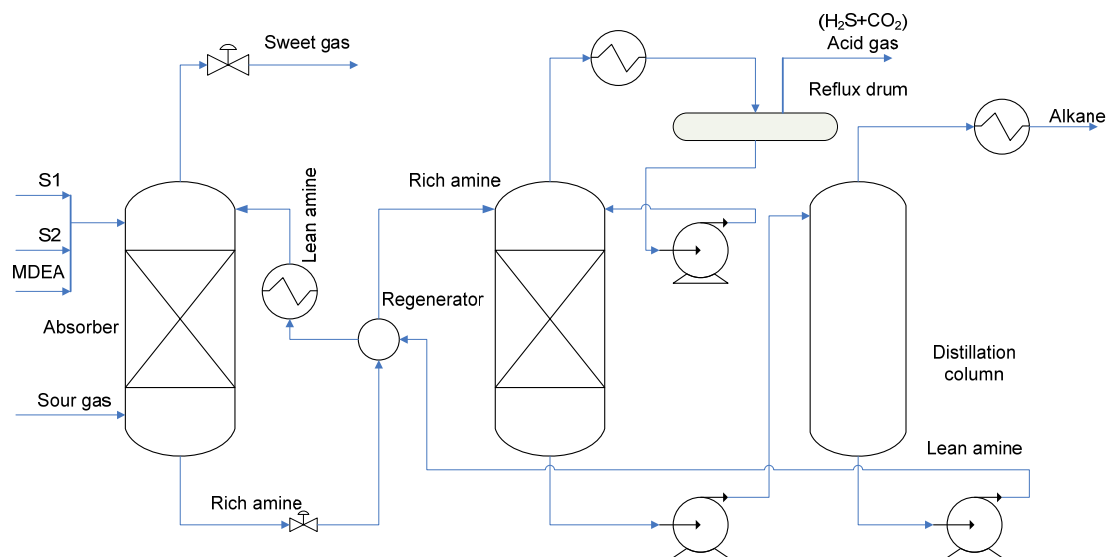


Figure 5.3. Acid Gas Removal Flowsheet

The objective of this design problem is to identify a solvent that can be used to reduce the consumption of MDEA by 50% and utilize all available recycle streams. The solvent when mixed with MDEA and the recycle streams must satisfy the environmental regulations. The solvent must also possess the qualities of efficient acid gas removal agents. Therefore, the molecular building blocks have to be selected such that the final structure should be similar to known efficient acid gas removal agents. The sink performance requirements are functions of vapor pressure (VP), heat of vaporization (H_v) and molar volume (V_m). Apart from the process constraints, the designed molecule should have minimum soil sorption coefficient ($\log K_{oc}$) to avoid accumulation of the escaping solvent in one place and a high toxic limit concentration (TLC) value. The energy index for the separation of the alkane from the final molecule must be low so that the alkane can be easily separated after the absorption.

Table 5.24: Property and Flowrate Data for Acid Gas Removal Problem

<i>Property</i>	<i>Feed Properties</i>		<i>MDEA Properties</i>	<i>Lower Bound for Sink</i>	<i>Upper Bound for Sink</i>
	<i>S₁</i>	<i>S₂</i>			
VP (mm Hg)	63.2	43.1	0.26	-	10
H _v (kJ/mol)	39	41	89	60	90
V _m (cm ³ /mol)	178	189	114	110	140
Flowrate (kmol/h)	50	70	180	300	350

5.4.2. Process Design

The first step in solving this integrated process and molecular design problem is to identify the targets for the molecular design from the process design constraints. The property operators corresponding to the target properties are defined by the following equations:

$$\psi_{VP} = \sum_{s=1}^{NS} x_s VP^{1.44} \quad (5.34)$$

$$\psi_{V_m} = \sum_{s=1}^{NS} x_s V_m \quad (5.35)$$

$$\psi_{H_v} = \sum_{s=1}^{NS} x_s H_v \quad (5.36)$$

The property targets for molecular design have been identified from eqs. (5.34)-(5.36).

They are listed in table 5.25.

Table 5.25: Property Targets for Molecular Design

<i>Property</i>	<i>Lower Bound</i>	<i>Upper Bound</i>
VP (mm Hg)	-	15.8
H _v (kJ/mol)	57	157
V _m (cm ³ /mol)	40	224

5.4.3. Molecular Design

Apart from the process constraints, there are two environmental constraints to be considered during the design of the new compounds. The maximum value of toxic limit concentration (*TLC*) is kept as 10 ppm and the soil sorption coefficient needs to be minimized. Now, the next step is to find suitable property models to predict these properties from molecular structure using topological indices or group contribution models.

For heat of vaporization, there is a reliable group contribution model available, which is given in eq. (5.37) (Marrero & Gani, 2001):

$$\Delta H_v = h_{v0} + \sum n_i h_{vi} \quad (5.37)$$

For vapor pressure, there are no group contribution relationships available. Nevertheless, there is an empirical relationship that can be used to calculate vapor pressure from boiling point (Sinha *et al.*, 2003). For boiling point, there is a group contribution expression available (Marrero & Gani, 2001):

$$\log VP = 5.58 - 2.7 \left(\frac{T_b}{T} \right)^{1.7} \quad (5.38)$$

$$T_b = t_{b0} \cdot \ln \left[\sum_{g=1}^{n_g} n_g t_{b1} + \sum_{s=1}^{N_s} n_s t_{b2} + \sum_{t=1}^{N_t} n_t t_{b3} \right] \quad (5.39)$$

Here, T_b is the boiling point of the liquid and T is the temperature at which VP is evaluated (323K in this example).

There are group contribution relationships available for the estimation of molar volume. However, a more accurate estimation technique is available for molar volume based on edge adjacency indices (Dai *et al.*, 1998). The available relationship is given in eq. (5.40):

$$V_m = 33.52\varepsilon + 30.67 \quad (5.40)$$

where ε is the edge adjacency index.

For toxic limit concentration, a QSAR model is available that uses connectivity index of order two (Koch, 1982):

$$\log(TLC) = 4.204 - 1.385 \left({}^2\chi^v \right) \quad (5.41)$$

Connectivity index of order two can be obtained from eq. (5.42)

$$({}^2\chi^v) = \sum_{i,j,k=1}^{p_2} [D(i)D(j)D(k)]^{-1/2} \quad (5.42)$$

where, p_2 is the number of all paths of length two in the molecular graph. $D(i)$, $D(j)$ and $D(k)$ are the valencies of vertices i , j and k , respectively.

For soil sorption coefficient, a QSAR model is available ($R=0.973$) that employs a variety of connectivity indices (Bahnick & Doucette, 1988):

$$\log(K_{oc}) = 0.53({}^1\chi) - 1.25(\Delta^1\chi^v) - 0.72(\Delta^0\chi^v) + 0.66 \quad (5.43)$$

where, $\Delta\chi$ is known as the delta connectivity index. It can be calculated using eq. (5.44):

$$(\Delta\chi) = (\chi)_{np} - \chi \quad (5.44)$$

where, χ_{np} is the molecular connectivity index of any height for the non-polar equivalent structure of the molecule. Thus, the topological indices involved in this case study are the following:

$${}^0\chi, {}^0\chi^v, {}^1\chi, {}^1\chi^v, {}^2\chi^v, \epsilon, \text{GC models}$$

The next step is to re-write the topological index based expressions in terms of signature descriptors. Equation (5.45) will provide the representation of the topological indices used in this case study and the signature descriptors:

$${}^0\chi = {}^1\alpha_G \cdot {}^0\chi(\text{root}({}^1\Sigma))$$

$$\begin{aligned}
{}^1\chi &= \frac{1}{2}({}^2\alpha_G) {}^1\chi(\text{root}({}^2\Sigma)) \\
\varepsilon &= \frac{1}{2}({}^3\alpha_i) \cdot \varepsilon(\text{root}({}^3\Sigma))
\end{aligned}
\tag{5.45}$$

$${}^2\chi = \frac{1}{2}({}^3\alpha_G) {}^2\chi(\text{root}({}^3\Sigma))$$

As discussed in chapter four, group contribution models with second order group contributions can be represented in terms of signatures of height two or three. The highest signature height among the topological indices is three as seen in eq. (5.45). Therefore, the maximum signatures height required in this problem is three. Now, the property operators are formed corresponding to the target properties and the upper and lower limits are calculated. The values are given in table 5.26.

Table 5.26: Property Operators and Targets

<i>Property</i>	Ω_j	<i>Lower Bound</i>	<i>Upper Bound</i>
<i>VP</i>	$\exp(T_b/t_{b0})$	6.75	-
<i>H_v</i>	$H_v - h_{v0}$	45.3	145.3
<i>V_m</i>	$(V_m - 30.67)/33.52$	0.28	5.75
<i>TLC</i>	$(4.204 - \log(TLC))/1.385$	2.21	-
$\log(K_{oc})$	$\log(K_{oc}) - 0.66$	Minimum	

In order to select the molecular building blocks, the structures of commonly used amine absorbents were identified:

Monoethanolamine	$\text{NH}_2\text{CH}_2\text{CH}_2\text{OH}$
Diethanolamine	$\text{OHCH}_2\text{CH}_2\text{NHCH}_2\text{CH}_2\text{OH}$
Methyl diethanolamine	$\text{OHCH}_2\text{CH}_2\text{N}(\text{CH}_3)\text{CH}_2\text{CH}_2\text{OH}$
Diisopropylamine	$(\text{CH}_3)_2\text{CHNHCH}(\text{CH}_3)_2$

All the molecular groups present in these molecules were selected as the building blocks for the new absorbent. In addition, to demonstrate of the ability of the signature-based algorithm to handle the design of molecules with multiple bonds, one additional group has been included:



Next, signatures are generated corresponding to the molecular groups. Only those signatures are selected, which form structures similar to the existing amine absorbents.

Now, an MILP problem can be formulated:

$$\text{Min } \Omega_{K_{oc}}$$

$$\Omega_{VP} \geq 6.75$$

$$\Omega_{H_v} \geq 45.3$$

$$\Omega_{H_v} \leq 145.3$$

$$\Omega_{V_m} \geq 0.28$$

$$\Omega_{V_m} \leq 5.75$$

$$\Omega_{TLC} \geq 2.21$$

$$\sum_i D_i x_i = 2 \left[\left(\sum_i x_i + \frac{1}{2} \sum_{\text{doublebonds}} x_i \right) - 1 \right]$$

(5.46)

$$\sum (l_i \rightarrow l_j)_h = \sum (l_j \rightarrow l_i)_h$$

$$\sum_{i=j} \eta_i x_i = 2K$$

$$\sum n_i x_i < \sum x_j$$

$$\Omega_{js} = \sum_{i=1}^N L_i x_i$$

$$\Omega_{js} = \sum_{i=1}^N C_i x_i + M \sum_{j=1}^N S_j x_j$$

The general structure of this set of equations is similar to the one in the previous case study. Different solutions are obtained by implementing integer cuts after each solution has been found. The best five solutions in terms of signatures are as follows:

Solution 1:

O1(C2(O1(C)C2(CC)))

O1(C2(O1(C)C2(NC)))

C1(N3(C1(N)C2(CN)C2(CN)))

C2(C2(C2(CC)O1(C)C2(N3(CCC)C2(CC)))

C2(O1(C2(OC))C2(C2(OC)C2(NC))

C2(O1(C2(OC))C2(C2(OC)N3(CCC))

C2(N3(C1(N)C2(NC)C2(NC))C2(C2(NC)C2(OC)))

C2(N3(C1(N)C2(NC)C2(NC))C2(C2(NC)O1(C)))

N3(C1(N3(CCC))C2(N3(CCC)C2(CC))C2(N3(CCC)C2(OC)))

Solution 2:

O1(C2(O1(C)C2(CC)))

O1(C2(O1(C)C2(NC)))
N2(C2(N2(CC)C2(OC))C2(N2(CC)C2(CC)))
C2(C2(C2(CC)C2(OC)C2(N2(CC)C2(CC)))
C2(C2(C2(CC)C2(NC)C2(O1(C)C2(CC)))
C2(C2(C2(CC)C2(NC)N2(C2(NC)C2(NC)))
C2(C2(O1(C)C2(NC)N2(C2(NC)C2(NC)))
C2(O1(C2(OC))C2(C2(OC)C2(CC))
C2(O1(C2(OC))C2(C2(OC)N2(CC))

Solution 3:

O1(C2(O1(C)C2(CC)))
O1(C2(O1(C)C2(NC)))
C1(N3(C1(N)C2(CN)C2(CN)))
C2(O1(C2(OC))C2(C2(OC)C3(=CC))
C2(O1(C2(OC))C2(C2(OC)N3(CCC))
C2(N3(C1(N)C2(NC)C2(NC))C2(C2(NC)C3(=CC)))
C2(N3(C1(N)C2(NC)C2(NC))C2(C2(NC)O1(C)))
N3(C1(N3(CCC))C2(N3(CCC)C2(CC))C2(N3(CCC)C2(OC)))
C2(C2(O1(C) C2(CC))C3(=C3(=CC)C2(CC)))
C2(C2(N3(CCC) C2(CC))C3(=C3(=CC)C2(CC)))
C3(=C3(=C3(=CC)C2(CC))C2(C3(=CC)C2(CO)))
C3(=C3(=C3(=CC)C2(CC))C2(C3(=CC)C2(NC)))

Solution 4:

O1(C2(O1(C)C2(NC)))

O1(C2(O1(C)C3(=CC)))

O1(C2(O1(C)C2(NC)))

C2(C2(C2(CC)C3(=CC)C2(N3(CCC)C2(CC)))

C2(C2(C2(CC)C2(NC)C3(=C3(=CC)C2(CC)))

C2(O1(C2(OC))C2(C2(OC)N3(CCC))

C2(O1(C2(OC))C3(C2(OC)C3(=CC))

C2(N3(C1(N)C2(NC)C2(NC))C2(C2(NC)C2(CC)))

C2(N3(C1(N)C2(NC)C2(NC))C2(C2(NC)O1(C)))

N3(C1(N3(CCC))C2(N3(CCC)C2(CC))C2(N3(CCC)C2(OC)))

C3(=C3(=C3(=CC)C2(CC))C2(C3(=CC)O1(C)))

C3(=C3(=C3(=CC)C2(OC))C2(C3(=CC)C2(CC)))

Solution 5:

O1(C2(O1(C)C2(CC)))

N1(C2(N1(C)C2(CC)))

C2(C2(C2(CC)C2(OC)C2(N1(C)C2(CC)))

C2(C2(C2(CC)C2(NC)C2(O1(C)C2(CC)))

C2(C2(C2(CC)C2(NC)N1(C2(NC)))

C2(O1(C2(OC))C2(C2(OC)C2(CC))

The final molecular structures are generated using the algorithm presented in section 4.7.7. A detailed description of the systematic generation of one solution using this procedure has been provided as well. The final molecular structures are given in figure 5.4.

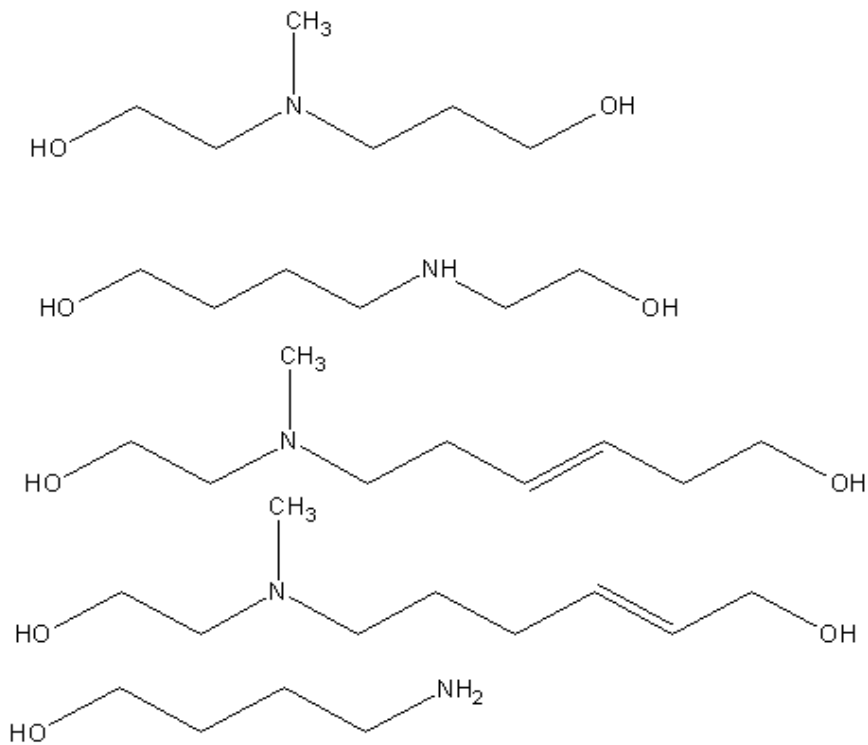


Figure 5.4. Best Five Solutions to Acid Gas Removal Problem

The input to the distillation column will be:

120 kmol/h alkane

90 kmol/h MDEA

90 kmol/h of the designed molecule

Next, the energy index needs to be estimated for each combination of the designed molecules with the alkane + MDEA mixture. However, at this time, the VLE data for the 2,5 dimethyl hexane - MDEA mixture is not readily available in the open literature. Therefore, this step is omitted for this problem and the candidates obtained from the molecular design can be considered as the close to final list.

5.4.4. Proof of Concept for Integrated Flowsheet and Molecular Design

Currently, the available data for solving integrated flowsheet and molecular design problems is very limited. Therefore, the immediate applicability of the developed integrated framework is limited to the design of systems that involve only relatively simple compounds. In order to illustrate the ability of the developed framework to link molecular design with flowsheet design, the previous case study is modified so that the necessary VLE data is available for the involved compounds. The following changes to the case study have been made to adapt the problem:

1. The alkane present in the recycle stream has been changed to cyclohexane.
2. The design of new molecules is limited to alkanes.
3. The property targets have been changed and the new targets are shown in table 5.27.

Table 5.27: New Property Targets for Acid Gas Absorbent Design

<i>Property</i>	<i>Lower Bound</i>	<i>Upper Bound</i>
<i>VP (mm Hg)</i>	-	600
<i>H_v (kJ/mol)</i>	30	100

V_m (cm ³ /mol)	40	224
TLC (ppm)	10	-

The problem is formulated analogously to the previous case study. However, the signatures have been defined only corresponding to alkane molecules. The best five solutions are:

Hexane

2-methyl hexane

3-methyl hexane

Heptane

2-methyl heptane

The next step is to estimate the energy index values for the "Cyclohexane – MDEA - Designed Molecule" mixtures. The VLE data for all the possible binary mixtures is available. The next step is to identify the maximum driving force for all the binary mixtures that could be encountered. The driving force based method for distillation column design employs eq. (2.34) (Bek-Pedersen & Gani, 2004):

$$D_{ij}(x_i, y_i, \alpha_{ij}) = \frac{x_i \alpha_{ij}}{1 + x_i (\alpha_{ij} - 1)} - x_i = y_i - x_i \quad (2.34)$$

The maximum driving force for each of the different binary mixtures is given in table 5.28.

Table 5.28: Driving Force Data

<i>Mixture</i>	<i>Driving Force</i>
Cyclohexane-MDEA	0.802
Cyclohexane-Hexane	0.0904
Hexane-MDEA	0.837
2-Methyl hexane-MDEA	0.782
2-Methyl hexane-Cyclohexane	0.0684
3-Methyl hexane-MDEA	0.775
3-Methyl hexane-Cyclohexane	0.0817
Heptane-MDEA	0.758
Heptane-Cyclohexane	0.1245
2-Methyl heptane-MDEA	0.6872
2-Methyl heptane-Cyclohexane	0.2581

Equation (2.35) is used to estimate the energy index values for all separations and the calculated values are given in table 5.29:

Table 5.29: Energy Index Values

<i>Molecule</i>	<i>Energy Index (GJ/h)</i>
2-Methyl heptane	59.3
Heptane	99.8
Hexane	129.4

3-Methyl hexane	140.8
2-Methyl hexane	164.4

It can be seen that the use of 2 Methyl heptane requires the least amount of energy compared to the rest of the molecules to be separated from the other compounds in the system. The lower energy index indicates potential savings related to separation, which makes this molecule the most promising candidate. This result is obvious from the driving forces of various binary mixtures given in table 5.28. However, the principles involved in this study can be extended into more challenging designs if the VLE data corresponding to the different mixtures is available. In more complicated designs, e.g. with more than three compounds, a number of flowsheets have to be identified and the energy index corresponding to all those structures must be estimated before a final solution can be determined. Once the input molecules are known, a software package (ICAS) developed at the Computer Aided Process Engineering Center (CAPEC) Denmark provides a computer implementation of the flowsheet design method.

6. Conclusions and Future Work

6.1. Major Achievements

The solution to an integrated process and product design problem is a challenging task for chemical engineers due to the complexity caused by the non-linear nature of the constitutive equations. Most of the current CAMD techniques use group contribution methods or algorithms based on connectivity indices. However, because of the non-linear nature of the expressions involved, it is difficult and sometimes impossible to achieve convergence unless some simplifying assumptions are employed. Because of this, most product design algorithms employ an iterative solution methodology or stochastic optimization techniques. However, these methods are computationally inefficient and obtaining a globally optimal solution is difficult. In addition, considering the process and product design problem separately can lead to suboptimal designs. In this work, a generalized approach has been developed for the simultaneous solution of both process and molecular design problems. A generalized algebraic approach has been developed for the solution of molecular synthesis problems. A systematic solution algorithm has been developed which can be applied in the solution of molecular synthesis problems with any number of property targets and with an unlimited number of molecular building blocks. Therefore, the limitations of earlier methods with respect to the number of properties and molecular building blocks have been alleviated. While a visual approach often

provides insights, the algebraic approach automatically generates the complete solution set which ensures that no potential solution is omitted or overlooked.

Another achievement is the development of higher order molecular property operators for use in the solution algorithm. The earlier models were using first order group contribution models. Because of the limitations of first order groups discussed earlier in this dissertation, the application of these algorithms was limited to the design of simple straight chain molecules with a limited number of carbon atoms. In this dissertation, molecular property operators have been developed that include second and third order group contribution methods. The logic followed in the development of these models utilizes that the higher order groups represent different interactions among the first order groups, and as such, it is possible to write expressions for those groups in terms of first order groups. Because of these higher order molecular groups, the algorithm can differentiate between structural isomers to some extent, especially in the design of structures with medium complexity. However, as the complexity of the molecules increases, the relative position of the groups becomes more important. Nonetheless, the developed methodology is capable of providing the basic structures of all possible candidate molecules. The developed algorithm is easily programmable and allows the designer to obtain an automatic solution set.

Furthermore, the methodology has been extended to include ring structures and aromatic compounds in the solution set. With the introduction of third order groups and the modifications to the solution algorithms, it is now even possible to design complicated structures like fused ring compounds and compounds with multiple rings.

A common problem faced during the design of molecules from a given set of molecular groups is that, some of the property contribution data may not be available for all of the groups being considered. It is also possible that some of the molecular groups are not identified in the group contribution (UNIFAC) tables. In such situations, the molecular operators cannot be formed because the missing property data from one group will alter the solution set. In this dissertation, a modified algorithm has been developed for incorporating the property contributions predicted by the combined group contribution-connectivity index technique (GC⁺ method) for cases when the property contributions of some of the candidate groups are not available in literature. Therefore, the algorithm developed for property based product design is even more general as it can be used for the design of molecules made from any set of molecular groups.

Property clustering is a relatively new technique and hence there are many potential areas where this concept can prove to be useful. The algorithms based on group contribution methods enabled the integration of process and product design. However, the applicability of these algorithms is limited to design based on properties for which group contribution models are available. Group contribution models are available for a limited number of properties and the accuracy of these methodologies for predicting properties other than thermodynamic properties can be questionable. In addition, for the design of molecules that need to satisfy many biological, environmental, health and safety property constraints, there is a need for more structural information than group contribution can provide. In the open literature, a variety of QSAR/QSPR relationships are available that make use of different topological indices. These relationships have been known in the product design field for a long time. However, the use of these models in

solving inverse problems was quite challenging because of the non-linear nature of most topological indices. The available techniques for such inverse problems include stochastic techniques based on genetic algorithms and Monte Carlo methods. There have been very few attempts to solve the inverse problem using a deterministic approach.

In this work, a new algorithm has been developed by introducing molecular signature descriptors into the reverse problem formulation framework. This approach has extended the applicability of the general RPF framework to include a wider range of property targets. The basic principles from graph theory are used to generate structural constraints that ensure meaningful solutions are obtained during the molecular design stage. The developed algorithm has the capability to incorporate a variety of QSAR/QSPR expressions based on different topological indices in inverse design. Apart from the topological index based expressions, group contribution based models can also be used in this approach. While designing the molecules, the contributions from the higher order molecular groups are also taken into consideration. One significant aspect of this new algorithm is that, its accuracy is not affected even if the design consists of compounds with heteroatoms, multiple bonds or ring structures. The prediction accuracy depends only upon on the accuracy of the original property models used in the forward problems.

Many QSAR/QSPR expressions make use of more than one topological index. In addition, different topological indices require molecular signatures of different heights in their transformation. In this work, methodologies have been developed to include signatures of different heights and obtain the solution in terms of the signatures with

maximum height. This transformation ensures that the solution obtained after the molecular design exhibits minimum degeneracy.

Similar to the previous algorithms, the solution to the molecular design problem is obtained as the number of appearances of the vectors that constitute the molecule. However, unlike for the molecular groups, identification of the bonds between different signatures is not a trivial issue. Therefore, an algorithm to generate molecular structures from the signatures has been developed as well. This algorithm can generate the solution with minimum degeneracy and no isomorphism among the solutions.

Finally, a novel framework to integrate process and product design with flowsheet design has been presented. This methodology will be helpful in screening out candidate molecules that make the process less efficient. Currently, flowsheet design models are available for the prediction of very few properties. However, current research in this field is expected to produce more models that can be incorporated into the existing framework. This will ensure that the process can be designed based on the molecule/molecules that yield the best performance in the most economical and safe manner.

6.2. Future Work

6.2.1. Statistical Tools for Product Design

Many properties are highly correlated and so, multivariate statistical methods can potentially be powerful tools in product design. Decomposition techniques like principal component analysis, partial least squares, factor analysis and cluster analysis can provide viable tools to develop more attribute-property relationships. Techniques have already been developed by Solvason *et al.* (2009) to directly use the regressors in the cluster

space. A similar approach for product design can provide more insights to include properties, which have previously not been considered in the design of molecules.

6.2.2. Integrated Process and Product Design

Recent developments in the area of simultaneous molecular and process design using group contribution methods (GCM) have been limited by models that are intended to predict properties at standard conditions. In most practical applications, the operating conditions do not correspond to the standard conditions. Furthermore, in the true spirit of simultaneous molecular and process design, the operating conditions should not be fixed a priori but determined during the simultaneous design procedure.

One of the practical problems with carrying out such an analysis is the difficulty in expressing dependence of the property as a function of the operating conditions. Therefore, the focus while solving such problems will be to bind the dependence on conditions on the boundary values rather than tracking the exact changes. To obtain a rigorous analysis of such functional dependences on properties, the analysis of the uncertainties in parameters should be done. The analysis of uncertainties that propagate through a model to affect the output will be a challenging problem (Enszer & Stadtherr, 2009). Still, for most processes the bounds on the process conditions may be available even if the exact probability distribution is difficult to obtain. In that case, the process can be modeled by interval analysis using a probability box (Enszer & Stadtherr, 2009). However, more work is needed in developing interval based modeling techniques.

When designing materials that will be used under challenging process conditions like high temperatures and pressures, the temporal effects must also be taken into

consideration. Dynamic modeling techniques should be incorporated in the next stage of research in this field to track property changes as a function of time.

The integration of flowsheet design with process and product design should be investigated further so that real industrial problems can be solved using the framework developed in this dissertation. The flowsheet design can be performed on a property platform using the driving force based method by d'Anterrosches and Gani (2005). In an integrated approach, the flowsheet properties will be estimated for each molecule that satisfies the property constraints. The optimum flowsheet structure will be finalized based on the molecules that result in a minimum value of the energy index. However, in addition to the energy index, a variety of other flowsheet properties are important from a process standpoint, e.g. safety index and risk index to name a few. Additionally, a variety of mass transfer based properties could possibly be directly linked to flowsheet properties. One example is the absorbent factor:

$$A_e = \frac{L}{KV} \quad (6.1)$$

Therefore, it is possible to generate relationships based on both process and molecular design parameters that affect flowsheet properties as well. The main focus should be to relate flowsheet properties to mass transfer based properties that will enable a more informed integrated design.

6.2.3. Inclusion of More Sophisticated Descriptors

The solution of molecular design problems has traditionally been considered as an iterative process (Harper *et al.*, 1999; Harper & Gani, 2000). Different classes of property models have been used in different stages of the overall process. The objective of this work will be to incorporate a different class of descriptors to eliminate the iterative steps in molecular design.

The molecular signature descriptor is a novel concept in reverse problem formulations. Therefore, its full potential is yet to be explored. The ability of signatures to form a variety of useful QSARs has already been established (Churchwell *et al.*, 2004; Weis *et al.*, 2005). However, very few QSAR relationships are formulated directly based on signatures. In all previous works, molecular signatures of a specific height have been used to generate the QSAR relationships. More qualitative analysis of the theoretical significance of signatures has to be performed. It has been shown that signatures of certain heights represent specific classes of topological indices. The theoretical interpretation of topological indices belonging to certain classes has been published (Randic & Zupan, 2001). Therefore, signatures of individual heights could also possess similar physical interpretations. Therefore, while generating QSAR relationships, these physical interpretations should also be analyzed. Based on the property for which signature based QSAR/QSPRs are generated, there will be signatures with more than one height. This will be synonymous with those QSAR/QSPR expressions that contain multiple topological indices belonging to different classes. Since the techniques incorporate signature descriptors in reverse problem formulations have already been

developed (Chemmgattuvalappil *et al.*, 2010), this work can also follow a similar approach.

An important limitation of current molecular design techniques is that the stability of the final molecular structures is not considered during the generation stage. This generally produces a large number of molecular structures with no physical meaning which have to be screened out afterwards. However, methodologies are available to calculate ring strain, which can be correlated to the stability of organic molecules (Benson, 1976). Apart from developing new QSAR/QSPR relationships and reverse engineering techniques, methodologies to ensure the stability of final structures should be developed, to make the existing algorithms more computationally efficient.

Until now, the reverse engineering techniques have been based on group contribution methods and to a lesser extent on two dimensional molecular descriptors. However, there are numerous three-dimensional molecular descriptors available to describe a variety of properties. It is important to bring these types of molecular descriptors into the reverse problem formulation framework. This would extend the state of reverse engineering to another level in a generalized multi-stage procedure for molecular design. The most significant application of this research is that, the molecular design will no longer be an iterative process (at least theoretically) and this work could significantly reduce the computational expense of solving molecular design problems.

6.2.4. Exploration of Biochemical Reaction Pathways

The exploration of new biochemical reaction pathways is a significant step in the design of sustainable chemical products. Many commodity chemicals produced from

petroleum-derived raw materials could possibly be produced in a much more sustainable manner by following novel yet unexplored reaction pathways (Broadbelt *et al.*, 2009). Similarly, new products could potentially be produced using a novel biochemical reaction pathway as well.

The exploration of new reaction pathways is possible through a network generation algorithm. The starting compounds and the expected final compounds are the input. There must be a collection of generalized enzyme functions, which could be based on existing data banks like KEGG (Kyoto Encyclopedia of Genes and Genomes) or CAS (Chemical Abstracts Service) registry. The fundamental principle is to generate a set of pathways or trees from raw material to product. The signature descriptors, which have the ability to represent changes in the species at each step as a change in the signature at the outermost level, could be employed to reduce the complexity of the reaction network generation algorithm.

The reaction network generation algorithm is analogous to molecular design algorithms, albeit more complicated, as no step in the reaction sequence is known at the time of design. The steps are automatically generated based on the constraints imposed by the specific enzymes involved and the target structures. A number of algorithms are available for the automatic generation of trees (Trinajstić, 1992). The signatures can be tailored to produce such trees representing the reaction sequences. Therefore, along with the pathways and compounds listed in databases, novel compounds and pathways could be generated in this manner. If the new product or pathway is more sustainable, it should replace the existing one. This work would be quite novel and required the tools include the signature descriptors and fundamental principles of graph theory.

6.2.5. Design of Reactions

Until now, the applications of reverse problem formulations have been limited to non-reactive systems. However, the new signature-based algorithms have the potential to extend the methodology to include reactive systems. For tracking the properties during a chemical reaction, the signature descriptor corresponding to a reaction can be defined as follows:

$${}^h\sigma(R) = f(\sigma(A)\sigma(B)) \quad (6.3)$$

where $\sigma(A)$ are the signatures corresponding to the reactants and $\sigma(B)$ are the signatures corresponding to the products involved in the reaction. The changes in the molecular groups during the reaction, will be reflected in the signatures. The signatures can be formed for both the reactants and the products. The relationships between reactants and products can be formed by considering them as un-connected graphs. A large databank corresponding to the different types of signatures will be required for applying this in actual industrial processes. For multi-stage reactions, there will be a change in signature at different stages. Since the signatures can be used to represent different topological indices, which are correlated to different properties, the changes in the signatures would provide a virtual representation of the changes in properties.

$$\Delta\theta = f({}^h\sigma(R)) \quad (6.4)$$

where $\Delta\theta$ is the change in the property function. The initial efforts in this research would be to develop functional relationships between the change in property functions and signatures. In this area of research, the principles of graph theory can be combined with the reaction mechanisms to identify the most favorable and feasible reactions. In a simultaneous process and product design problem, the changes in the process conditions will also have an effect on the properties. Therefore, the tools discussed earlier in this chapter would have to be included in this work. This research would enable the design of reactive systems purely on a property platform.

References

1. Achenie, L. E. K., Gani, R., & Venkatasubramanian, V. (2003). *Computer Aided Molecular Design: Theory and practice*: Elsevier.
2. Achenie, L. E. K., & Sinha, M. (2004). The design of blanket wash solvents with environmental considerations. *Advances in Environmental Research*, 8, 213-227.
3. Ambrose, D. (1978). Correlation and estimation of vapor-liquid critical properties: I. Critical temperatures of organic compounds. *NPL Report Chem* (United Kingdom, National Physical Laboratory, Division of Chemical Standards), 92, 35
4. Ambrose, D. (1980). Vapor-liquid critical properties. *NPL Report Chem* (United Kingdom, National Physical Laboratory, Division of Chemical Standards), 107, 61
5. Bahnick, D. A., & Doucette, W. J. (1988). Use of molecular connectivity indexes to estimate soil sorption coefficients for organic chemicals. *Chemosphere*, 17, 1703-1715.
6. Balaban, A. T. (2001). *QSPR/QSAR Studies by Molecular descriptors*. New york: Nova Science Publishers, Inc.
7. Baskin, I. I., Gordeeva, E. V., Devdariani, R. O., Zefirov, N. S., Palyulin, V. A., & Stankevich, M. I. (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.
8. Bek-Pedersen, E., & Gani, R. (2004). Design and synthesis of distillation systems using a driving-force-based approach. *Chemical Engineering and Processing*, 43, 251.

9. Benson, S. W. (1968). Thermo chemical kinetics, Methods for the Estimation of thermochemical Data and Rate Parameters. New york: Wiley.
10. Benson, S. W. (1976). Thermochemical Kinetics: Methods for the estimation of thermochemical data and rate parameters. New york: John Wiley & Sons.
11. Bondy, J. A., & Murty, U. S. R. (2008). Graph Theory: Springer.
12. Broadbelt, L. J., Henry, C. S., & Hatzimanikatis, V. (2009). Discovery of novel routes for the production of fuels and chemicals. Design for Energy and Environment.
13. Cabezas, H. (2000). Designing green solvents. Chemical Engineering, 107-109.
14. Camarda, K. V., & Maranas, C. D. (1999). Optimization in Polymer Design Using Connectivity Indices. Industrial & Engineering Chemistry Research, 38, 1884-1892.
15. Camarda, K. V., & Sunderesan, P. (2005). An Optimization Approach to the Design of Value-Added Soybean Oil Products. Industrial & Engineering Chemistry Research, 44, 4361.
16. Chemmangattuvalappil, N. G., Eljack, F. T., Solvason, C. C., & Eden, M. R. (2009a). A novel algorithm for molecular synthesis using enhanced property operators. Computers & Chemical Engineering, 33, 636-643.
17. Chemmangattuvalappil, N. G., Solvason, C. C., Bommareddy, S., & Eden, M. R. Reverse problem formulation approach to molecular design using property operators based on signature descriptors. Computers & Chemical Engineering, 34, 2062-2071.
18. Chemmangattuvalappil, N. G., Solvason, C. C., Bommareddy, S., & Eden, M. R. (2009b). Incorporating Molecular Signature Descriptors in Reverse Problem Formulations. Computer Aided Chemical Engineering, 27, 73-78.

19. Chemmangattuvalappil, N. G., Solvason, C. C., Bommareddy, S., & Eden, M. R. (2009c). Novel Molecular Design Technique Using Property Operators Based on Signature Descriptors. *Computer Aided Chemical Engineering*, 27, 897-902.
20. Chemmangattuvalappil, N. G., Solvason, C. C., Bommareddy, S., & Eden, M. R. (2009d). A Systematic Methodology for Molecular Synthesis using Combined Property Clustering and GC+ Methods. *Design for Energy and the Environment*, 757-766.
21. Chemmangattuvalappil, N. G., Solvason, C. C., Bommareddy, S., & Eden, M. R. (2010a). Combined property clustering and GC⁺ techniques for process and product design. *Computers & Chemical Engineering*, 34, 582-591.
22. Chemmangattuvalappil, N. G., Solvason, C. C., Bommareddy, S., & Eden, M. R. (2010b). Molecular Signature Descriptors for Integrated Flowsheet and Molecular Design. *Computer Aided Chemical Engineering*, 1267-1272.
23. Chemmangattuvalappil, N. G., Solvason, C. C., & Eden, M. R. (2009e). Property based product design using combined property clustering and GC+ techniques. *Computer-Aided Chemical Engineering*, 26, 237-242.
24. Chemmangattuvalappil, N. G., Solvason, C. C., Eljack, F.T., & Eden, M. R. (2008). Enhanced Algebraic Property Clustering Technique for Molecular Synthesis. *Computer Aided Chemical Engineering*, 25.
25. Churchwell, C. J., Rintoul, M. D., Martin, S., Visco, D. P., Kotu, A., Larson, R. S., Sillerud, L. O., Brown, D. C., & Faulon, J.-L. (2004). The signature molecular descriptor 3. Inverse-quantitative structure-activity relationship of ICAM-1 inhibitory peptides. *Journal of Molecular Graphics & Modelling*, 22, 263-273.

26. Constantinou, L., & Gani, R. (1994). New group contribution method for estimating properties of pure compounds. *AIChE Journal*, 40, 1697-1710.
27. Constantinou, L., Gani, R., & O'Connell, J. P. (1995). Estimation of the acentric factor and the liquid molar volume at 298 K using a new group contribution method. *Fluid Phase Equilibria*, 103, 11.
28. Constantinou, L., Prickett, S. E., & Mavrovouniotis, M. L. (1993). Estimation of thermodynamic and physical properties of acyclic hydrocarbons using the ABC approach and conjugation operators. *Industrial & Engineering Chemistry Research*, 32, 1734-1746.
29. Conte, E., Martinho, A., Matos, H. A., & Gani, R. (2008). Combined Group-Contribution and Atom Connectivity Index-Based Methods for Estimation of Surface Tension and Viscosity. *Industrial & Engineering Chemistry Research*, 47, 7940-7954.
30. Contreras, M. L., Rozas, R., & Valdivia, R. (1994). Exhaustive Generation of Organic Isomers. 3. Acyclic, Cyclic, and Mixed Compounds. *Journal of Chemical Information and Computer Sciences*, 34, 610.
31. Cornell, J.A. (2002). *Experiments with Mixtures*. New York: John Wiley & Sons.
32. Cox, D. R. (1971). A note on polynomial response functions for mixtures. *Biometrika*, 58, 155-159.
33. Cussler, E. L., & Moggridge, G. D. (2001). *Chemical Product Design*: Cambridge University Press.
34. Cussler, E. L., Wagner, A., & Marchal-Heussler, L. (2010). Designing chemical products requires more knowledge of perception. *AIChE Journal*, 56, 283-288.
35. d'Anterrosches, L., & Gani, R. (2005). Group contribution based process flowsheet synthesis, design and modelling. *Fluid Phase Equilibria*, 228-229, 141-146.

36. Dai, J., Jin, L., & Wang, L. (1998). Prediction of molar volume of aliphatic compounds using edge adjacency index. *Progress in Natural Science*, 8, 754-761.
37. Davidson, S. (2002). Fast Generation of an Alkane-Series Dictionary Ordered by Side-Chain Complexity. *Journal of Chemical Information and Computer Sciences*, 42, 147.
38. Eden, M. R., Jorgensen, S. B., Gani, R., & El-Halwagi, M. (2002). Property integration - a new approach for simultaneous solution of process and molecular design problems. *Computer Aided Chemical Engineering*, 10, 79-84.
39. Eden, M. R., Jorgensen, S. B., Gani, R., & El-Halwagi, M. M. (2004). A novel framework for simultaneous separation process and product design. *Chemical Engineering and Processing*, 43, 595-608.
40. El-Halwagi, M. M., Glasgow, I. M., Qin, X., & Eden, M. R. (2004). Property integration: Componentless design techniques and visualization tools. *AIChE Journal*, 50, 1854-1869.
41. Eljack, F. T., Abdelhady, A. F., Eden, M. R., Gabriel, F. B., Qin, X., & El-Halwagi, M. M. (2005). Targeting optimum resource allocation using reverse problem formulations and property clustering techniques. *Computers & Chemical Engineering*, 29, 2304-2317.
42. Eljack, F. T., Solvason, C.C. & Eden, M.R. (2007a). An Algebraic Property Clustering Technique for Molecular Design. *Computer Aided Chemical Engineering (CD Volume)*.

43. Eljack, F. T., & Eden, M. R. (2008). A systematic visual approach to molecular design via property clusters and group contribution methods. *Computers & Chemical Engineering*, 32, 3002-3010.
44. Eljack, F. T., Eden, M. R., Kazantzi, V., Qin, X., & El-Halwagi, M. M. (2007b). Simultaneous process and molecular design-A property based approach. *AIChE Journal*, 53, 1232-1239.
45. Enszer, J. A., & Stadtherr, M. A. (2009). Rigorous propagation of imprecise probabilities in process models. *Design for Energy and Environment*, 77-92.
46. Eslick, J. C., Ye, Q., Park, J., Topp, E. M., Spencer, P., & Camarda, K. V. (2009). A computational molecular design framework for crosslinked polymer networks. *Computers & Chemical Engineering*, 33, 954.
47. Estrada, E. (1995a). Edge Adjacency Relationships and a Novel Topological Index Related to Molecular Volume. *Journal of Chemical Information and Computer Sciences*, 35, 31-33.
48. Estrada, E. (1995b). Edge adjacency relationships in molecular graphs containing heteroatoms: a new topological index related to molar volume. *Journal of Chemical Information and Computer Sciences*, 35, 701-707.
49. 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, 731-740.
50. Faulon, J.-L., Churchwell, C. J., & Visco, D. P., Jr. (2003a). The Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences. *Journal of Chemical Information and Computer Sciences*, 43, 721-734.

51. Faulon, J.-L., Visco, D. P., Jr., & Pophale, R. S. (2003b). The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *Journal of Chemical Information and Computer Sciences*, 43, 707-720.
52. Fedors, R. F. (1982). A relationship between chemical structure and the critical temperature. *Chemical Engineering Communications*, 16, 149-151.
53. Gani, R. (2004a). Chemical product design: challenges and opportunities. *Computers & Chemical Engineering*, 28, 2441-2457.
54. Gani, R. (2004b). Computer-aided methods and tools for chemical product design. *Chemical Engineering Research and Design*, 82, 1494-1504.
55. Gani, R., & Constantinou, L. (1996). Molecular structure based estimation of properties for process design. *Fluid Phase Equilibria*, 116, 75-86.
56. Gani, R., Harper, P. M., & Hostrup, M. (2005). Automatic Creation of Missing Groups through Connectivity Index for Pure-Component Property Prediction. *Industrial & Engineering Chemistry Research*, 44, 7262-7269.
57. Gani, R., & O'Connell, J. P. (2001). Properties and CAPE: from present uses to future challenges. *Computers & Chemical Engineering*, 25, 3-14.
58. Gani, R., & Pistikopoulos, E. N. (2002). Property modelling and simulation for product and process design. *Fluid Phase Equilibria*, 194-197, 43-59.
59. Garg, S., & Achenie, L. E. K. (2001). Mathematical Programming Assisted Drug Design for Nonclassical Antifolates. *Biotechnology Progress*, 17, 412-418.
60. Ghosh, P., Sundaram, A., Venkatasubramanian, V., & Caruthers, J. M. (2000). Integrated product engineering: a hybrid evolutionary framework. *Computers & Chemical Engineering*, 24, 685.

61. Gordeeva, E. V., Molchanova, M. S., & Zefirov, N. S. (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, 389-415.
62. Gutman, I., & Polansky, O. E. (1986). *Mathematical Concepts in Organic Chemistry*. Berlin: Springer-verlag.
63. Hall, L. H., Dailey, R. S., & Kier, L. B. (1993a). 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, 598.
64. Hall, L. H., Kier, L. B., & Frazer, J. W. (1993b). 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, 148.
65. Harper, P. M., & Gani, R. (2000). A multi-step and multi-level approach for computer aided molecular design. *Computers & Chemical Engineering*, 24, 677-683.
66. Harper, P. M., Gani, R., Kolar, P., & Ishikawa, T. (1999). Computer-aided molecular design with combined molecular modeling and group contribution. *Fluid Phase Equilibria*, 158-160, 337-347.
67. Hill, M. (2009). Chemical Product Engineering - The third paradigm. *Computers & Chemical Engineering*, 33, 947-953.
68. Horvath, A. L. (1992). Molecular Design. Chemical Structure Generation from the Properties of Pure Organic Compounds. [In: *Stud. Phys. Theor. Chem.*, 1992; 75].

69. 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, 2332-2339.
70. Joback, K. G., & Reid, R. C. (1987). Estimation of pure-component properties from group contributions. *Chemical Engineering Communications*, 57, 233-243.
71. Joback, K. G., & Stephanopoulos, G. (1989). Designing molecules possessing desired physical property values. In *Proceedings FOCAPD89*. Austin, TX: CACHE corporation.
72. Katritzky, A. R., & Gordeeva, E. V. (1993). Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *Journal of Chemical Information and Computer Sciences*, 33, 835-857.
73. Kazantzi, V., Qin, X., El-Halwagi, M., Eljack, F., & Eden, M. (2007). Simultaneous Process and Molecular Design through Property Clustering Techniques: A Visualization Tool. *Industrial & Engineering Chemistry Research*, 46, 3400-3409.
74. Kehiaian, H. V. (1983). Group contribution methods for liquid mixtures: a critical review. *Fluid Phase Equilibria*, 13, 243-252.
75. Kier, L. B., & Hall, L. H. (1976). Molecular connectivity VII. Specific treatment of heteroatoms. *Journal of Pharmaceutical Sciences*, 65, 1806-1809.
76. Kier, L. B., & Hall, L. H. (1986). *Chemometrics Series, 9: Molecular Connectivity in Structure-Activity Analysis*. New York: John Wiley & Sons.
77. Kier, L. B., Hall, L. H., & Frazer, J. W. (1993a). 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, 143-147.

78. Kier, L. B., Hall, L. H., & Frazer, J. W. (1993b). 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, 143.

79. Koch, R. (1982). Molecular connectivity and acute toxicity of environmental pollutants. *Chemosphere*, 11, 925-931.

80. Kokossis, A. C., & Yang, Y. (2009). Future system challenges in the design of renewable bio-energy systems and the systems of sustainable biorefineries. *Design for Energy and Environment*.

81. Kvasnicka, V., & Pospichal, J. (1990). Canonical indexing and constructive enumeration of molecular graphs. *Journal of Chemical Information and Computer Sciences*, 30, 99-105.

82. Kvasnicka, V., & Pospichal, J. (1996). Simulated Annealing Construction of Molecular Graphs with Required Properties. *Journal of Chemical Information and Computer Sciences*, 36, 516-526.

83. Lin, B., Chavali, S., Camarda, K., & Miller, D. C. (2005). Computer-aided molecular design using Tabu search. *Computers & Chemical Engineering*, 29, 337.

84. Marcoulaki, E. C., & Kokossis, A. C. (1998). Molecular design synthesis using stochastic optimisation as a tool for scoping and screening. *Computers & Chemical Engineering*, 22, S11.

85. Marrero, J., & Gani, R. (2001). Group-contribution based estimation of pure component properties. *Fluid Phase Equilibria*, 183-184, 183.
86. Martin, T. M., & Young, D. M. (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, 1378-1385.
87. Mavrovouniotis, M. L. (1990). Estimation of properties from conjugate forms of molecular structures: the ABC approach. *Industrial & Engineering Chemistry Research*, 29, 1943-1953.
88. McLeese, S. E., Eslick, J. C., Hoffmann, N. J., Scurto, A. M., & Camarda, K. V. (2010). Design of ionic liquids via computational molecular design. *Computers & Chemical Engineering*, In Press, Corrected Proof.
89. Montgomery, D. C. (2005). *Design and Analysis of Experiments*. USA: John Wiley & Sons.
90. Murray, W. J., Hall, L. H., & Kier, L. B. (1975). Molecular connectivity. III: Relationship to partition coefficients. *Journal of Pharmaceutical Sciences*, 64, 1978-1981.
91. Ortiz, P. J., & Perez, C. S. (1982). *Elementos de mecanica cuantica y estructura atomica*: Universidad de La Babana.
92. Papadopoulos, A. I., & Linke, P. (2006). Efficient integration of optimal solvent and process design using molecular clustering. *Chemical Engineering Science*, 61, 6316.
93. Qin, X., Gabriel, F., Harell, D., & El-Halwagi, M. M. (2004). Algebraic Techniques for Property Integration via Componentless Design. *Industrial & Engineering Chemistry Research*, 43, 3792-3798.

94. R.Daudel, R.Leferbvre, & C.Moser. (1959). Quantum chemistry, methods and applications. New york: Interscience Publishers, Inc.
95. Raman, V. S., & Maranas, C. D. (1998). Optimization in product design with properties correlated with topological indices. *Computers & Chemical Engineering*, 22, 747.
96. Randic, M. (1975). Characterization of molecular branching. *Journal of the American Chemical Society*, 97, 6609-6615.
97. Randic, M., & Basak, S. C. (2001). A new descriptor for structure-property and structure-activity correlations. *Journal of Chemical Information and Computer Sciences*, 41, 650-656.
98. Randic, M., & Zupan, J. (2001). On Interpretation of Well-Known Topological Indices. *Journal of Chemical Information and Computer Sciences*, 41, 550-560.
99. Reid, R. C., Prausnitz, J. M., & Pauling, B. E. (1987). *The properties of gases and liquids* (4 ed.). New York: McGraw-Hill.
100. Russel, B. M., Henriksen, J. P., Jørgensen, S. B., & Gani, R. (2000). Integration of design and control through model analysis. *Computers & Chemical Engineering*, 24, 967.
101. Sahinidis, N. V., Tawarmalani, M., & Yu, M. (2003). Design of alternative refrigerants via global optimization. *AIChE Journal*, 49, 1761-1775.
102. Shelley, M. D., & El-Halwagi, M. M. (2000). Component-less design of recovery and allocation systems: a functionality-based clustering approach. *Computers & Chemical Engineering*, 24, 2081.

103. Sheridan, R. P., & Kearsley, S. K. (1995). Using a Genetic Algorithm To Suggest Combinatorial Libraries. *Journal of Chemical Information and Computer Sciences*, 35, 310-320.
104. Siddhaye, S., Camarda, K., Southard, M., & Topp, E. (2004). Pharmaceutical product design using combinatorial optimization. *Computers & Chemical Engineering*, 28, 425.
105. Sinha, M., & Achenie, L. E. K. (2001). Systematic design of blanket wash solvents with recovery considerations. *Advances in Environmental Research*, 5, 239-249.
106. Sinha, M., Achenie, L. E. K., & Gani, R. (2003). Blanket wash solvent blend design using interval analysis. *Industrial & Engineering Chemistry Research*, 42, 516-527.
107. Skvortsova, M. I., Baskin, I. I., Slovokhotova, O. L., Palyulin, V. A., & Zefirov, N. S. (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, 630-634.
108. Smith, B. V., & Ierapetritou, M. (2009). Framework for Consumer-Integrated Optimal Product Design. *Industrial & Engineering Chemistry Research*, 48, 8566.
109. Solvason, C. C., Chemmangattuvalappil, N. G., Eljack, F. T., & Eden, M. R. (2009). Efficient Visual Mixture Design of Experiments using Property Clustering Techniques. *Industrial & Engineering Chemistry Research*, 48, 2245-2256.
110. Solvason, C. C., Eljack, F. T., Chemmangattuvalappil, N. G., & Eden, M. R. (2008). Visual Mixture Design Using Property Clustering. *Computer Aided Chemical Engineering*, 25.

111. Stephanopoulos, G. (2003). Invention and innovation in a Product-Centered Chemical Industry: General trends and a case study. In AIChE. San Fransisco, CA.
112. Trinajstic, N. (1992). Chemical Graph Theory (2nd Edition ed.). Boca Raton, FL: CRC press.
113. Uchida, M. (1980). Affinity and mobility of fungicidal dialkyl dithiolanylidenemalonates in rice plants. *Pesticide Biochemistry and Physiology*, 14, 249.
114. Venkatasubramanian, V., Chan, K., & Caruthers, J. M. (1994). Computer-aided molecular design using genetic algorithms. *Computers & Chemical Engineering*, 18, 833-844.
115. Venkatasubramanian, V., Chan, K., & Caruthers, J. M. (1995). Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm. *Journal of Chemical Information and Computer Sciences*, 35, 188-195.
116. Visco, D. P., Pophale, R. S., Rintoul, M. D., & Faulon, J.-L. (2002). Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor. *Journal of Molecular Graphics & Modelling*, 20, 429-438.
117. Weis, D. C., Faulon, J.-L., LeBorne, R. C., & Visco, D. P., Jr. (2005). The Signature Molecular Descriptor. 5. The Design of Hydrofluoroether Foam Blowing Agents Using Inverse-QSAR. *Industrial & Engineering Chemistry Research*, 44, 8883-8891.
118. 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, 2636-2638.

119. Wilson, R. J. (1986). Introduction to Graph Theory: Longman Scientific & Technical.
120. Wu, H. S., & Sandler, S. I. (1989). Proximity effects on the predictions of the UNIFAC model: I. Ethers. *AIChE Journal*, 35, 168-172.
121. Wu, H. S., & Sandler, S. I. (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, 881-889.

Appendix: Group Contribution and Connectivity Indices Data

The property contributions used in the case studies are given in Tables A.1 through A.6. In the first three tables, three levels of property contributions given in the property model used by Marrero & Gani (2001) are given. The following properties are predicted using these models:

- Normal boiling and melting temperatures
- Critical pressure, critical volumes and critical temperature
- Standard enthalpy of vaporization and standard Gibbs energy, and standard enthalpy of formation

Table A.4 and Table A.5 provide the group contribution values estimated in the property models used by (Constantinou *et al.*, 1995) for the prediction of molar volume and acentric factor.

Table A.6 provides the regressed parameters in the equation for the connectivity index method. These expressions are used for the prediction of properties predicted by Marrero & Gani (2001) models, but typically used for predicting the contributions of the missing groups.

Table A.1: First Order Group Contribution Property Data (Marrero & Gani, 2001)

First-order groups and their contributions along with sample assignments

Group	Example	T_{mli}	T_{bli}	T_{cli}	P_{cli}	V_{cli}	G_{fli}	H_{fli}	H_{vli}	H_{lusti}	
1	CH ₃	<i>n</i> -Tetracontane (2)	0.6953	0.8491	1.7506	0.018615	68.35	2.878	-42.479	0.217	1.660
2	CH ₂	<i>n</i> -Tetracontane (38)	0.2515	0.7141	1.3327	0.013547	56.28	8.064	-20.829	4.910	2.639
3	CH	2-Methylpentane (1)	-0.3730	0.2925	0.5960	0.007259	37.50	8.254	-7.122	7.962	0.134
4	C	2,2-Dimethylbutane (1)	0.0256	-0.0671	0.0306	0.001219	16.01	16.413	8.928	10.730	-1.232
5	CH ₂ =CH	1-Hexene (1)	1.1728	1.5596	3.2295	0.025745	111.43	95.738	57.509	4.031	1.268
6	CH=CH	2-Hexene (1)	0.9460	1.5597	3.0741	0.023003	98.43	92.656	69.664	9.456	4.441
7	CH ₂ =C	2-Methyl-1-butene (1)	0.7662	1.3621	2.7717	0.021137	91.40	85.107	61.625	8.602	2.451
8	CH=C	2-Methyl-2-butene (1)	0.1732	1.2971	2.5666	0.019609	83.89	88.691	81.835	14.095	3.032
9	C=C	2,3-Dimethyl-2-butene (1)	0.3928	1.2739	2.6391	0.014114	90.66	93.119	95.710	19.910	2.616
10	CH ₂ =C=CH	1,2-Butadiene (1)	1.7036	2.6840	5.4330	0.035483	143.57	229.906	198.840	11.310	7.076
11	CH ₂ =C=C	3-Methyl-1,2-butadiene (1)	1.5453	2.4014	4.8219	0.029678	146.36	226.710	208.490	*****	7.435
12	CH=C=CH	2,3-Pentadiene (1)	1.2850	2.5400	*****	*****	*****	*****	*****	*****	6.000
13	CH≡C	1-Pentyne (1)	2.2276	1.7618	3.7897	0.014010	84.60	230.029	224.902	6.144	-1.548
14	C≡C	3-Decyne (1)	2.0516	1.6767	4.5870	0.010888	74.66	216.013	228.282	12.540	6.128
15	aCH	Benzene (6)	0.5860	0.8365	2.0337	0.007260	42.39	26.732	12.861	3.683	1.948
16	aC fused with aromatic ring	Naphthalene (2)	1.8955	1.7324	5.4979	0.003564	35.71	20.379	20.187	6.631	0.845
17	aC fused with non-aromatic subring	Indane (2)	1.2065	1.1995	3.1058	0.006512	34.65	33.912	30.768	6.152	1.095
18	aC except as above	Benzophenone (1)	0.9176	1.5468	4.5344	0.012859	26.47	23.331	24.701	6.824	-0.531
19	aN in aromatic ring	Pyridine (1)	2.0438	1.3977	4.0954	-0.003339	36.47	89.902	70.862	9.420	2.555
20	aC-CH ₃	Toluene (1)	1.0068	1.5653	3.4611	0.020907	97.33	24.919	-19.258	8.279	2.969
21	aC-CH ₂	Ethylbenzene (1)	0.1065	1.4925	2.9003	0.018082	87.19	31.663	4.380	11.981	0.948
22	aC-CH	Cumene (1)	-0.5197	0.8665	1.9512	0.011795	73.51	30.393	18.440	13.519	-1.037
23	aC-C	<i>tert</i> -Butylbenzene (1)	-0.1041	0.5229	0.8576	0.011298	67.20	40.127	35.297	16.912	-2.856
24	aC-CH=CH ₂	Styrene (1)	1.2832	2.4308	5.7861	0.030637	134.69	114.531	77.863	*****	4.013
25	aC-CH=CH	1-Propenylbenzene (1)	1.7744	2.9262	6.5062	0.026282	128.84	111.216	88.084	*****	8.274
26	aC-C=CH ₂	α -Methylstyrene (1)	1.2612	2.1472	4.9967	0.026371	110.74	115.728	90.927	*****	3.324
27	aC-C≡CH	Phenylacetylene (1)	1.7495	2.3057	6.4572	0.019507	112.08	263.205	257.448	*****	2.514
28	aC-C=C	1-Phenyl-1-propyne (1)	*****	2.7341	*****	*****	*****	*****	*****	*****	*****
29	OH	1,4-Butanediol (2)	2.7888	2.5670	5.2188	-0.005401	30.61	-144.051	-178.360	24.214	4.786
30	aC-OH	Phenol (1)	5.1473	3.3205	9.3472	-0.008788	50.77	-131.327	-164.191	34.099	8.427
31	COOH	1,5-Pentanedioic acid (2)	7.4042	5.1108	14.6038	0.009885	90.66	-337.090	-389.931	17.002	10.692
32	aC-COOH	Benzoic acid (1)	12.4296	6.0677	15.4515	0.017100	119.10	-312.422	-361.249	*****	14.649
33	CH ₃ CO	2-Butanone (1)	2.9588	3.1178	7.0058	0.025227	127.99	-120.667	-180.604	15.195	8.062
34	CH ₂ CO	3-Pentanone (1)	2.5232	2.6761	5.7157	0.019619	112.79	-120.425	-163.090	19.392	8.826
35	CHCO	2,4-Dimethyl-3-pentanone (1)	1.1565	2.1748	4.4743	0.012487	97.16	-116.799	-139.909	20.350	7.205
36	CCO	2,2,4,4-Tetramethyl-3-pentanone (1)	1.0638	1.7287	*****	*****	*****	*****	*****	*****	*****
37	aC-CO	Acetophenone (1)	2.9157	3.4650	9.4806	0.011007	90.69	-91.812	-106.965	25.036	4.852
38	CHO	1-Hexanal (1)	3.0186	2.5388	5.8013	0.010204	71.08	-100.882	-130.816	12.370	11.325
39	aC-CHO	Benzaldehyde (1)	2.4744	3.5172	9.4795	0.019633	122.91	-80.222	-107.159	*****	7.273
40	CH ₃ COO	Butyl acetate (1)	2.1657	3.1228	6.3179	0.033812	148.91	-306.733	-387.458	19.342	7.910
41	CH ₂ COO	Methyl Butyrate (1)	1.6329	2.9850	5.9619	0.026983	132.89	-298.332	-364.204	21.100	9.479
42	CHCOO	Ethyl isobutyrate (1)	1.0668	2.2869	4.7558	0.021990	125.52	-301.414	-352.057	24.937	9.317
43	CCOO	Ethyl 2,2-dimethylpropionate (1)	0.3983	1.6918	*****	*****	*****	*****	*****	23.739	*****
44	HCOO	Propyl formate (1)	2.0223	2.5972	5.6064	0.015249	93.29	-276.878	-327.678	15.422	8.115

45	aC-COO	Methyl benzoate (1)	1.3348	3.1952	6.7311	0.018948	105.53	-291.662	-307.727	25.206	8.149
46	aC-OOCH	Phenyl formate (1)	*****	0.4621	*****	*****	*****	*****	*****	*****	*****
47	aC-OOC	Phenyl acetate (1)	4.8044	3.0854	*****	*****	*****	*****	*****	*****	5.875
48	COO except as above	Ethyl acrylate (1)	1.5038	2.1903	4.7346	0.013087	81.17	-299.803	-331.397	*****	10.573
49	CH ₃ O	Methyl butyl ether (1)	1.3643	1.7703	3.4393	0.020084	88.20	-90.329	-156.062	5.783	5.089
50	CH ₂ O	Di- <i>n</i> -butyl ether (1)	0.8733	1.3368	2.4217	0.017954	74.03	-105.579	-152.239	9.997	4.891
51	CH-O	sec-Butyl ether (1)	0.2461	0.8924	0.7889	0.014487	60.06	-101.207	-147.709	14.620	4.766
52	C-O	<i>tert</i> -Butylether (1)	-0.4446	0.4983	0.2511	0.005613	52.96	-92.804	-121.608	13.850	2.458
53	aC-O	Methyl phenyl ether (1)	1.3045	1.8522	3.6588	0.005115	47.27	-83.354	-101.783	16.151	-0.118
54	CH ₂ NH ₂	Ethylamine (1)	3.2742	2.7987	8.1745	0.011413	117.62	68.812	-10.703	15.432	13.482
55	CHNH ₂	sec-Butylamine (1)	30.8394	2.0948	4.2847	0.013049	76.36	61.452	0.730	16.048	6.283
56	CNH ₂	<i>tert</i> -Butylamine (1)	11.7400	1.6525	2.8546	0.010790	80.01	55.202	2.019	17.257	*****
57	CH ₃ NH	Dimethylamine (1)	2.4034	2.2514	4.5529	0.015863	77.04	88.512	24.740	11.831	4.490
58	CH ₂ NH	Dipropylamine (1)	1.7746	1.8750	3.2422	0.020482	95.15	88.874	23.610	13.067	7.711
59	CHNH	Diisopropylamine (1)	1.7577	1.2317	2.0057	0.005329	99.16	73.101	21.491	14.048	2.561
60	CH ₃ N	Methyldiethylamine (1)	0.9607	1.3841	3.0106	0.021186	94.94	125.906	55.024	9.493	6.008
61	CH ₂ N	Triethylamine (1)	0.0442	1.1222	2.1673	0.027454	74.05	121.247	65.331	12.636	1.756
62	aC-NH ₂	Aniline (1)	3.9889	3.8298	10.2155	0.005335	81.40	66.470	17.501	23.335	6.542
63	aC-NH	<i>N</i> -methyl aniline (1)	1.4837	2.9230	8.4081	-0.005596	86.37	98.195	53.274	23.026	0.624
64	aC-N	<i>N,N</i> -dimethyl aniline (1)	1.7618	2.1918	5.8536	-0.000838	108.39	143.280	115.606	22.249	-2.576
65	NH ₂ except as above	Cyclobutylamine	3.3478	2.0315	4.7420	0.000571	63.39	42.687	-8.556	13.425	6.158
66	CH=N	Acetaldazine (2)	8.8492	1.5332	*****	*****	*****	*****	*****	*****	*****
67	C=N	Ketazine (2)	1.4621	1.4291	*****	*****	*****	*****	*****	*****	*****
68	CH ₂ CN	Propionitrile (1)	2.5760	4.5871	12.9827	0.036523	133.62	134.997	99.245	21.923	7.303
69	CHCN	Isobutyronitrile (1)	2.1393	3.9774	8.4309	0.029034	134.73	142.475	151.390	24.963	9.464
70	CCN	2,2-Dimethylpropionitrile (1)	3.3807	2.8870	5.8829	0.024654	120.74	142.295	124.770	24.967	4.166
71	aC-CN	Benzonitrile (1)	5.1346	4.1424	10.4124	0.020978	119.08	162.175	148.968	*****	6.788
72	CN except as above	Acrylonitrile (1)	3.2747	3.0972	8.1381	0.024346	94.91	130.986	124.917	16.639	6.867
73	CH ₂ NCO	Ethyl isocyanate (1)	4.2256	3.4891	*****	*****	*****	*****	*****	*****	*****
74	CHNCO	Isopropyl isocyanate (1)	*****	3.1220	*****	*****	*****	*****	*****	*****	*****
75	CNCO	<i>tert</i> -Butyl isocyanate (1)	9.1492	*****	*****	*****	*****	*****	*****	*****	*****
76	aC-NCO	Phenyl isocyanate (1)	2.2327	3.1853	6.5884	0.025065	141.24	*****	*****	*****	*****
77	CH ₂ NO ₂	1-Nitropropane (1)	3.2131	4.5311	10.9507	0.021056	157.57	25.783	-65.620	29.640	10.989
78	CHNO ₂	2-Nitropropane (1)	0.7812	3.8069	9.5487	0.014899	143.36	16.407	-60.750	29.173	*****
79	CNO ₂	2-Methyl-2-nitropropane (1)	5.6280	3.3059	*****	*****	*****	*****	*****	*****	-4.187
80	aC-NO ₂	Nitrobenzene (1)	4.3531	4.5750	12.1243	0.018311	133.06	57.352	-22.931	24.863	7.572
81	NO ₂ except as above	Nitrocyclohexane (1)	3.0376	3.2069	*****	*****	*****	*****	*****	*****	6.302
82	ONO	Butyl nitrite (1)	*****	1.8896	*****	*****	*****	*****	*****	*****	*****
83	ONO ₂	<i>n</i> -Butyl nitrate (1)	2.5974	3.2656	*****	*****	*****	*****	*****	*****	9.353
84	HCON(CH ₂) ₂	Diethylformamide (1)	*****	5.8779	*****	*****	*****	*****	*****	*****	*****
85	HCONHCH ₂	Ethylformamide (1)	*****	7.4566	*****	*****	*****	*****	*****	46.490	*****
86	CONH ₂	Butyramide (1)	13.2124	6.5652	25.1184	0.001467	138.71	-127.512	-201.369	44.240	16.840
87	CONHCH ₃	Methylacetamide (1)	5.4720	5.0724	20.5590	0.023455	190.71	-102.912	-203.069	*****	17.429
88	CONHCH ₂	Ethylacetamide (1)	5.8825	6.6810	*****	*****	*****	*****	-183.613	52.723	*****
89	CON(CH ₃) ₂	Dimethylacetamide (1)	4.1720	6.0070	15.4603	0.043090	244.71	-56.412	-188.069	38.290	11.553
90	CONCH ₂ CH ₂	Methylethylacetamide (1)	*****	*****	*****	*****	*****	*****	-48.210	*****	*****
91	CON(CH ₂) ₂	Diethylacetamide (1)	*****	5.0664	*****	*****	*****	*****	*****	*****	*****
92	CONHCO	Diacetamide (1)	9.1763	7.6172	*****	*****	*****	*****	*****	*****	*****
93	CONCO	Methyldiacetamide	3.2657	5.6487	*****	*****	*****	*****	*****	*****	*****

Group	Example	T_{mLi}	T_{bLi}	T_{cLi}	P_{cLi}	V_{cLi}	G_{cLi}	H_{Li}	H_{vLi}	H_{fusLi}	
94	aC-CONH ₂	Benzamide	12.8071	8.3775	*****	*****	*****	*****	*****	16.811	
95	aC-NH(CO)H	<i>N</i> -phenylformamide (1)	5.6631	7.3497	19.8979	0.023447	162.08	-44.595	-125.052	8.658	
96	aC-N(CO)H	<i>N</i> -methyl- <i>N</i> -phenylmethanamide (1)	3.3602	5.1373	*****	*****	*****	*****	*****	*****	
97	aC-CONH	<i>N</i> -methylbenzamide (1)	6.5160	7.5850	*****	*****	*****	*****	*****	10.959	
98	aC-NHCO	<i>N</i> -(2-methylphenyl)acetamide (1)	9.8204	7.4955	*****	*****	*****	*****	*****	4.370	
99	aC-NCO	Phenylmethylacetamide (1)	7.2552	*****	*****	*****	*****	*****	*****	*****	
100	NHCONH	<i>N,N'</i> -dimethylurea (1)	9.3110	8.9406	*****	*****	*****	*****	*****	9.862	
101	NH ₂ CONH	Methylurea (1)	14.2020	16.3539	*****	*****	*****	*****	*****	12.845	
102	NH ₂ CON	<i>N,N</i> -dimethylurea (1)	13.0856	2.0796	*****	*****	*****	*****	*****	10.958	
103	NHCON	Trimethylurea (1)	8.4447	7.1529	*****	*****	*****	*****	*****	12.098	
104	NCON	Tetramethylurea (1)	3.5041	4.1459	*****	*****	*****	*****	*****	9.557	
105	aC-NHCONH ₂	Phenylurea (1)	13.4695	5.7604	*****	*****	*****	*****	*****	16.703	
106	aC-NHCONH	<i>N,N'</i> -diphenylurea	23.2570	1.1633	*****	*****	*****	*****	*****	18.460	
107	NHCO except as above	<i>N</i> -chloroacetamide (1)	3.0882	*****	*****	*****	*****	*****	*****	*****	
108	CH ₂ Cl	1-Chlorobutane (1)	1.9253	2.6364	6.2561	0.021419	112.12	-19.484	-65.056	11.754	6.353
109	CHCl	2-Chloropropane (1)	1.0224	2.0246	4.3756	0.015640	100.78	-31.933	-65.127	12.048	*****
110	CCl	2-Chloro-2-methylpropane (1)	1.8424	1.7049	3.7063	0.009187	87.01	-37.848	-62.881	16.597	-0.082
111	CHCl ₂	1,1-Dichloroethane (1)	2.5196	3.3420	7.8956	0.028236	159.79	-24.214	-80.812	17.251	6.781
112	CCl ₂	2,2-Dichloropropane (1)	3.6491	2.9609	*****	*****	*****	*****	*****	20.473	1.823
113	CCl ₃	1,1,1-Trichloroethane	4.4493	3.9093	8.8073	0.036746	204.71	-44.122	-105.369	20.550	3.492
114	CHF ₂	1-Fluorobutane (1)	1.5597	1.5022	3.3179	0.023315	87.71	-180.212	-227.469	8.238	7.139
115	CHF	2-Fluorobutane (1)	1.1289	1.3738	2.6702	0.020040	78.08	-228.239	-261.901	*****	3.917
116	CF	2-Fluoro-2-methylpropane (1)	2.5398	1.0084	2.1633	-0.010120	*****	*****	*****	6.739	*****
117	CHF ₂	1,1-Difluoroethane (1)	2.1689	2.2238	3.5702	0.031524	102.71	-411.239	-463.901	*****	7.011
118	CF ₂	Perfluorohexane (4)	0.1312	0.5142	0.8543	0.018572	95.09	*****	*****	1.621	*****
119	CF ₃	Hexafluoroethane (2)	1.4828	1.1916	1.7737	0.048565	108.85	-615.333	-673.875	7.352	2.526
120	CCl ₂ F	Tetrachloro-1,2-difluoroethane (2)	3.2035	2.5053	5.1653	0.037948	171.04	-249.020	-306.765	8.630	3.114
121	HCClF	1-Chloro-1,2,2,2-tetrafluoroethane (1)	*****	2.0542	*****	*****	*****	*****	*****	*****	*****
122	CClF ₂	1,2-Dichlorotetrafluoroethane (2)	1.7510	1.7227	3.0593	0.041641	146.01	-396.814	-458.074	8.086	2.156
123	aC-Cl	Chlorobenzene (1)	1.7134	2.0669	5.7046	0.016033	92.67	1.985	-17.002	11.224	4.435
124	aC-F	Hexafluorobenzene (6)	0.9782	0.7945	1.5491	0.014037	54.36	-141.306	-160.965	3.965	2.003
125	aC-I	Iodobenzene (1)	2.1905	3.7739	12.4470	0.014403	131.08	91.505	95.048	*****	2.814
126	aC-Br	Bromobenzene (1)	2.4741	2.8414	8.4199	0.010199	104.12	42.977	38.917	14.393	5.734
127	I ⁻ except as above	Iodoethane (1)	1.9444	3.1778	8.5775	-0.004637	104.28	43.910	47.632	14.171	6.103
128	Br ⁻ except as above	Bromoethane (1)	1.7641	2.4231	4.5036	-0.001460	77.99	5.528	-1.703	9.888	4.826
129	F ⁻ except as above	Benzyl fluoride (1)	1.2308	0.8504	0.8976	0.012034	24.62	-182.973	-201.968	*****	3.096
130	Cl ⁻ except as above	Ethyl chloroacetate (1)	1.5454	1.5147	4.0947	0.007923	57.77	-29.876	-46.963	*****	5.181
131	CHNOH	Propionaldehyde oxime (1)	3.9813	4.5721	*****	*****	*****	*****	*****	*****	*****
132	CNOH	Diethyl ketoxime (1)	3.5484	4.0142	*****	*****	*****	*****	*****	*****	*****
133	aC-CHNOH	Phenyl oxime (1)	10.5579	*****	*****	*****	*****	*****	*****	*****	*****
134	OCH ₂ CH ₂ OH	2-Ethoxyethanol (1)	2.3651	4.8721	10.4579	0.025986	159.33	-233.335	-343.903	31.493	8.454
135	OCHCH ₂ OH	2-Ethoxy-1-propanol (1)	*****	4.2329	*****	*****	*****	*****	*****	*****	*****
136	OCH ₂ CHOH	1-Methoxy-2-propanol (1)	1.5791	3.6653	*****	0.018783	147.66	-239.423	-333.385	*****	12.594
137	-O-OH	<i>tert</i> -Butylhydroperoxide (1)	4.8181	3.1669	5.8307	-0.002815	58.01	-75.568	-125.111	*****	*****
138	CH ₂ SH	Ethaneethiol (1)	2.2992	3.1974	7.7300	0.017299	105.68	27.469	-8.021	16.815	10.068
139	CHSH	2-Propanethiol (1)	0.9704	2.5910	5.8527	0.008968	109.36	27.030	3.510	17.098	4.266
140	CSH	2-Methyl-2-propanethiol (1)	4.2329	2.0902	4.6431	0.005118	94.01	27.338	12.589	18.397	-0.623

141	aC-SH	Benzenethiol (1)	2.8464	3.2675	9.5115	0.010086	95.08	48.905	41.648	17.413	4.513
142	-SH (except as above)	Cyclohexanethiol (1)	0.9600	2.3323	7.7987	0.006399	57.89	15.818	11.339	9.813	5.829
143	CH ₃ S	Dimethylsulfide (1)	1.7150	2.9892	6.9733	0.018013	122.03	35.845	-3.337	14.296	7.497
144	CH ₂ S	Diethylsulfide (1)	1.0063	2.6524	6.4871	0.015254	106.60	42.684	21.492	16.965	4.096
145	CHS	Diisopropylsulfide (1)	0.7892	2.0965	*****	*****	*****	*****	*****	19.038	*****
146	CS	di- <i>tert</i> -Butylsulfide (1)	1.1170	1.6412	*****	*****	*****	*****	*****	19.996	*****
147	aC-S-	Phenyl methyl sulfide (1)	0.9646	2.9731	*****	*****	*****	*****	*****	*****	*****
148	SO	Dimethyl sulfoxide (1)	5.3663	6.2796	19.8953	-0.005534	82.36	-52.231	-71.050	*****	13.403
149	SO ₂	Dimethyl sulfone (1)	7.0778	7.0976	17.2586	-0.000784	89.95	-257.608	-305.498	*****	17.748
150	SO ₃ (sulfite)	Dimethyl sulfite (1)	*****	3.9199	8.6910	0.004240	115.80	*****	-430.833	*****	*****
151	SO ₃ (sulfonate)	Dimethyl sulfonate (1)	5.8426	6.7785	*****	*****	*****	*****	*****	*****	*****
152	SO ₄ (sulfate)	Dimethyl sulfate (1)	3.6976	5.5627	18.9366	-0.027208	144.58	-519.853	-621.412	*****	*****
153	aC-SO	Phenyl methyl sulfoxide (1)	3.9911	6.1185	*****	*****	*****	*****	*****	*****	*****
154	aC-SO ₂	Diphenyl sulfone (1)	5.2948	8.4333	*****	*****	135.47	-314.643	-370.493	*****	3.281
155	PH (phosphine)	Dimethylphosphine (1)	*****	2.0536	*****	*****	*****	*****	*****	*****	*****
156	P (phosphine)	Trimethylphosphine (1)	*****	1.0984	*****	*****	*****	*****	*****	*****	*****
157	PO ₃ (phosphite)	Triethylphosphite (1)	1.0306	2.7900	*****	*****	*****	*****	*****	*****	*****
158	PHO ₃ (phosphonate)	Dimethylphosphonate (1)	*****	5.6433	*****	*****	*****	*****	*****	*****	*****
159	PO ₃ (phosphonate)	Trimethylphosphonate (1)	*****	4.5468	*****	*****	*****	*****	*****	*****	*****
160	PHO ₄ (phosphate)	Diethylphosphate (1)	2.7461	5.1567	*****	*****	*****	*****	*****	*****	*****
161	PO ₄ (phosphate)	Trimethylphosphate (1)	2.0330	3.7657	16.9914	-0.029036	85.59	*****	-1060.325	*****	*****
162	aC-PO ₄	Triphenylphosphate (1)	-1.7840	2.3522	*****	*****	*****	*****	-1005.161	*****	4.256
163	aC-P	Triphenylphosphine (1)	0.2337	2.9272	38.6148	-0.126108	-142.79	*****	72.339	*****	-5.654
164	CO ₃ (carbonate)	Diethylcarbonate (1)	3.6593	2.8847	6.6804	0.007235	93.56	-447.186	-516.282	21.613	8.363
165	C ₂ H ₃ O	Ethyl oxirane (1)	1.3135	2.8451	6.6418	0.021238	125.43	11.149	-52.241	*****	*****
166	C ₂ H ₂ O	2,2-Dimethyl oxirane (1)	*****	2.6124	6.0159	0.010678	194.36	1.890	-51.390	*****	*****
167	C ₂ O	Trimethyl oxirane (1)	*****	2.2036	*****	*****	*****	*****	*****	*****	*****
168	CH ₂ (cyclic)	Cyclopentane (5)	0.5699	0.8234	1.8815	0.009884	49.24	13.287	-18.575	3.341	1.069
169	CH (cyclic)	Methylcyclopentane (1)	0.0335	0.5946	1.1020	0.007596	44.95	6.107	-12.464	6.416	2.511
170	C (cyclic)	1,1-Dimethylcyclohexane (1)	0.1695	0.0386	-0.2399	0.003268	33.32	-0.193	-2.098	7.017	-0.921
171	CH=CH (cyclic)	Cyclobutene (1)	1.1936	1.5985	3.6426	0.013815	83.91	86.493	59.841	7.767	1.185
172	CH=C (cyclic)	1-Methylcyclopentene (1)	0.4344	1.2529	3.5475	0.010576	70.98	67.056	64.295	7.171	2.559
173	C=C (cyclic)	1,2-Dimethylcyclopentene (1)	0.3048	1.1975	*****	*****	*****	*****	*****	*****	*****
174	CH ₂ =C (cyclic)	Methylene cyclohexane (1)	0.2220	1.5109	4.4913	0.019101	83.96	*****	*****	*****	5.351
175	NH (cyclic)	Cyclopentimine (1)	3.4814	2.1634	5.9726	-0.003678	51.80	72.540	23.138	13.700	8.655
176	N (cyclic)	<i>N</i> -methylpyrrolidine (1)	0.6040	1.6541	4.3905	-0.001179	31.41	83.779	65.622	*****	0.269
177	CH=N (cyclic)	Imidazole (1)	5.5779	6.5230	*****	*****	*****	*****	*****	*****	3.993
178	C=N (cyclic)	2-Methyl-1H-imidazole (1)	6.6382	6.6710	*****	*****	*****	*****	*****	*****	*****
179	O (cyclic)	Tetrahydropyran (1)	1.3828	1.0245	2.7409	-0.000387	17.69	-114.062	-137.353	6.877	3.806
180	CO (cyclic)	Cyclobutanone (1)	3.2119	2.8793	12.6396	-0.000207	57.38	-156.672	-180.166	17.124	6.137
181	S (cyclic)	2-Methyl-thiophene (1)	1.6023	2.3256	5.5523	0.001540	45.45	12.020	15.453	12.262	5.170
182	SO ₂ (cyclic)	Cyclobutadiene sulfone (1)	6.1006	*****	24.3995	0.002487	96.66	-241.601	-283.839	*****	9.934

Table A.2: Second Order Group Contribution Property Data (Marrero & Gani, 2001)

Second-order groups and their contributions along with sample assignments

Group	Example	T_{m2j}	T_{b2j}	T_{c2j}	P_{c2j}	V_{c2j}	G_{t2j}	H_{t2j}	H_{v2j}	H_{us2j}	
1	$(CH_3)_2CH$	2-Methylpentane (1)	0.1175	-0.0035	-0.0471	0.000473	1.71	-0.418	-0.419	-0.399	0.396
2	$(CH_3)_3C$	2,2,4,4-Tetramethylpentane (2)	-0.1214	0.0072	-0.1778	0.000340	3.14	-2.776	-1.967	-0.417	0.554
3	$CH(CH_3)CH(CH_3)_2$	2,3,4-Trimethylpentane (2)	0.2390	0.3160	0.5602	-0.003207	-3.75	6.996	6.065	0.532	-1.766
4	$CH(CH_3)C(CH_3)_2$	2,2,3,4,4-Pentamethylpentane (2)	-0.3276	0.3976	0.8994	-0.008733	-10.06	8.938	8.078	0.623	0.351
5	$C(CH_3)_2C(CH_3)_2$	2,2,3,3,4,4-Hexamethylpentane (2)	3.3297	0.4487	1.5535	-0.016852	-8.70	10.735	10.535	5.086	-1.089
6	$CH_n=CH_m-CH_p=CH_k$ (k, m, n, p in 0..2)	1,3-Butadiene (1)	0.7451	0.1097	0.4214	0.000792	-7.88	-6.562	-11.786	1.632	1.408
7	$CH_3-CH_n=CH_o$ (m, n in 0..2)	2-Methyl-2-butene (3)	0.0524	0.0369	-0.0172	-0.000101	0.50	-0.120	-0.048	0.064	0.070
8	$CH_2-CH_m=CH_n$ (m, n in 0..2)	1,4-Pentadiene (2)	-0.1077	-0.0537	0.0262	0.000815	0.14	1.006	1.449	-0.060	-0.632
9	$CH_p-CH_m=CH_n$ (m, n in 0..2; p in 0..1)	3-Methyl-1-butene (1)	-0.2485	-0.0093	-0.1526	-0.000163	-2.67	3.857	3.964	0.004	-0.368
10	CHCHO or CCHO	2-Methylbutyraldehyde (1)	0.5715	-0.1286	-1.0434	0.005789	10.36	-0.525	1.514	-0.550	-0.369
11	CH_3COCH_2	2-Pentanone (1)	-0.0968	-0.0215	-0.0338	-0.000111	-4.08	-1.543	0.033	-0.403	0.105
12	CH_3COCH or CH_3COC	3-Methyl-2-pentanone (1)	-0.6024	-0.0803	-0.3658	-0.001892	3.02	2.202	4.994	0.723	1.005
13	CHCOOH or CCOOH	2-Methyl butanoic acid (1)	-3.1734	-0.3203	-4.7275	0.006916	10.56	3.920	1.121	7.422	5.475
14	CH_3COOCH or CH_3COOC	Isopropyl acetate (1)	0.2114	-0.2066	-0.5537	-0.000569	4.28	-11.779	-12.295	-1.871	1.208
15	CO-O-CO	Propanoic anhydride (1)	-1.2441	-0.0500	-0.3576	0.001812	2.98	-16.075	-14.140	*****	-2.666
16	CHOH	2-Butanol (1)	-0.3489	-0.2825	-0.6768	0.000246	-3.04	-5.614	-4.422	-0.206	-0.599
17	COH	2-Methyl-2-butanol (1)	0.3695	-0.5325	-1.5224	0.003224	13.98	-25.382	-25.929	-1.579	-0.459
18	CH_3COCH_nOH (n in 0..2)	3-Hydroxy-2-butanone (1)	0.9886	-0.2987	-0.3940	-0.002912	5.17	6.621	8.244	*****	*****
19	NCCOH or NCCOH	2-Hydroxypropionitrile (1)	-1.1810	0.2981	0.3414	-0.000516	0.68	4.833	0.000	*****	-0.149
20	$OH-CH_n-COO$ (n in 0..2)	Ethyl lactate (1)	-0.1526	-0.2310	*****	*****	*****	*****	*****	*****	*****
21	$CH_m(OH)CH_n(OH)$ (m, n in 0..2)	Ethylene glycol (1)	-0.0414	0.8854	1.9395	-0.004712	7.54	-1.051	-0.592	-6.611	-0.306
22	$CH_m(OH)CH_n(-)$ (m, n, p in 0..2)	2-Amino-1-butanol (1)	-0.5941	0.5082	1.2342	0.002581	5.58	-1.506	-0.959	*****	-0.041
23	$CH_m(NH_2)CH_n(NH_2)$ (m, n in 0..2)	Ethylenediamine (1)	0.3258	-0.0064	-3.3555	0.000726	20.82	0.344	-1.443	2.384	-1.575
24	$CH_m(NH)CH_n(NH_2)$ (m, n in 1..2)	Diethylenetriamine (1)	-1.8403	0.2318	-1.1598	0.000157	-26.31	3.848	3.608	*****	*****
25	$H_2NCOCH_nCH_mCONH_2$ (m, n in 1..2)	Butanediamide (1)	11.5351	*****	*****	*****	*****	*****	*****	*****	*****
26	$CH_m(NH_n)-COOH$ (m, n in 0..2)	1-Alanine (1)	12.3481	*****	62.4740	-0.002696	17.78	3.145	6.598	*****	7.032
27	$HOOC-CH_n-COOH$ (n in 1..2)	Malonic acid (1)	0.9327	-0.1222	1.9595	-0.001479	12.46	-5.217	-6.058	*****	4.264
28	$HOOC-CH_n-CH_m-COOH$ (n, m in 1..2)	Succinic acid (1)	7.5057	*****	0.7686	0.000090	15.17	-4.281	-6.929	*****	29.245
29	$HO-CH_n-COOH$ (n in 1..2)	2-Hydroxyisobutyric acid (1)	-0.4531	-0.4625	*****	*****	*****	*****	*****	*****	*****
30	$NH_2-CH_n-CH_m-COOH$ (n, m in 1..2)	β -Alanine (1)	14.1593	*****	*****	*****	*****	*****	*****	*****	*****
31	CH_3-O-CH_n-COOH (n in 1..2)	Methoxyacetic acid (1)	-2.3026	0.9198	0.4750	-0.001445	7.91	-2.678	-1.727	*****	*****
32	HS-CH-COOH	2-Mercaptopropionic acid (1)	-2.1535	*****	*****	*****	*****	*****	*****	*****	*****
33	$HS-CH_n-CH_m-COOH$ (n, m in 1..2)	β -Thiolactic acid (1)	-2.7514	*****	-0.2697	0.000655	20.43	-7.376	7.292	*****	-3.623
34	$NC-CH_n-CH_m-CN$ (n, m in 1..2)	1,2-Dicyanoethane (1)	4.0747	1.8957	1.9699	0.002330	24.82	18.974	5.661	*****	-8.038
35	$OH-CH_n-CH_m-CN$ (n, m in 1..2)	3-Hydroxypropanenitrile (1)	-0.9493	1.3434	0.2311	-0.001022	14.54	0.558	-3.906	*****	-4.371
36	$HS-CH_n-CH_m-SH$ (n, m in 1..2)	1,2-Ethanedithiol (1)	0.2232	0.1815	2.1272	0.001321	-10.31	6.728	0.794	-0.683	-0.931
37	$COO-CH_n-CH_m-OOC$ (n, m in 1..2)	Ethylene glycol diacetate (1)	-0.5946	0.3401	1.5418	-0.003385	-2.33	1.306	4.025	1.203	*****
38	$OOC-CH_m-CH_n-COO$ (n, m in 1..2)	Dimethylsuccinate (1)	2.5962	0.5794	*****	*****	*****	*****	*****	*****	2.303
39	$NC-CH_n-COO$ (n in 1..2)	Methylcyanoacetate (1)	-0.2509	1.2171	2.7051	-0.001999	-0.73	*****	*****	*****	1.100
40	$COCH_nCOO$ (n in 1..2)	Methylacetoacetate (1)	0.6304	0.2427	0.7502	-0.000231	1.69	10.556	-7.261	*****	*****
41	$CH_m-O-CH_n=CH_p$ (m, n, p in 0..3)	Ethyl vinyl ether (1)	-0.0811	0.1399	0.2900	-0.000432	-4.54	-10.098	-9.411	0.372	3.169
42	$CH_m=CH_n-F$ (m, n in 0..2)	1-Fluoro-1-propene (1)	-0.2568	0.0591	*****	*****	*****	*****	*****	*****	2.823
43	$CH_m=CH_n-Br$ (m, n in 0..2)	1-Bromo-1-propene (1)	-0.4329	-0.3192	*****	-0.010021	2.63	14.470	17.014	*****	2.212
44	$CH_m=CH_n-I$ (m, n in 0..2)	1-Iodo-1-propene (1)	*****	-0.3486	*****	*****	*****	*****	*****	*****	*****
45	$CH_m=CH_n-Cl$ (m, n in 0..2)	1-Chloro-2-methylpropene (1)	0.0446	-0.0268	-0.0188	0.000152	2.80	8.207	9.715	*****	-0.480
46	$CH_m=CH_n-CN$ (m, n in 0..2)	Acrylonitrile (1)	0.1027	0.0653	-1.1249	0.000893	3.82	-8.304	-16.903	*****	-0.405

47	$\text{CH}_n=\text{CH}_m-\text{COO}-\text{CH}_p$ (m, n, p in 0..3)	Ethyl Acrylate (1)	0.2117	-0.0430	-0.0880	0.000044	0.21	-12.085	-12.509	*****	-0.014
48	$\text{CH}_m=\text{CH}_n-\text{CHO}$ (m, n in 0..2)	Propenaldehyde (1)	-0.7191	0.1102	*****	*****	*****	*****	*****	*****	*****
49	$\text{CH}_m=\text{CH}_n-\text{COOH}$ (m, n in 0..2)	Acrylic Acid (1)	2.4103	0.0667	-1.7762	-0.000763	4.36	10.194	9.090	*****	1.291
50	$\text{aC}-\text{CH}_n-\text{X}$ (n in 1..2) X: Halogen	Benzyl bromide (1)	0.8092	0.4537	2.2630	0.002464	-4.88	-8.081	-8.570	*****	*****
51	$\text{aC}-\text{CH}_n-\text{NH}_m$ (n in 1..2; m in 0..2)	Benzyl amine (1)	-1.0802	0.2590	1.4069	-0.000034	2.50	-2.044	-3.447	4.608	-0.639
52	$\text{aC}-\text{CH}_n-\text{O}-$ (n in 1..2)	Benzyl ethyl ether (1)	0.8607	-0.0425	0.2698	-0.000417	-7.49	6.043	5.486	*****	0.969
53	$\text{aC}-\text{CH}_n-\text{OH}$ (n in 1..2)	Benzyl alcohol (1)	0.8981	0.1005	-1.0107	0.002944	-0.25	*****	*****	*****	-2.754
54	$\text{aC}-\text{CH}_n-\text{CN}$ (n in 1..2)	Benzyl cyanide (1)	0.1088	1.0587	2.4950	-0.000796	-11.01	25.157	16.950	*****	*****
55	$\text{aC}-\text{CH}_n-\text{CHO}$ (n in 1..2)	Phenyl acetaldehyde (1)	1.9470	-0.0177	*****	*****	*****	*****	*****	*****	*****
56	$\text{aC}-\text{CH}_n-\text{SH}$ (n in 1..2)	Phenyl methanethiol (1)	1.2057	0.1702	0.8705	0.000183	2.00	16.725	7.568	*****	0.890
57	$\text{aC}-\text{CH}_n-\text{COOH}$ (n in 1..2)	Phenyl acetic acid (1)	0.3666	0.1584	*****	*****	*****	*****	*****	*****	-4.086
58	$\text{aC}-\text{CH}_n-\text{CO}-$ (n in 1..2)	Phenyl acetone (1)	-0.2363	0.3094	*****	*****	*****	*****	*****	*****	*****
59	$\text{aC}-\text{CH}_n-\text{S}-$ (n in 1..2)	Benzyl methyl sulfide (1)	0.4506	0.1030	*****	*****	*****	*****	*****	*****	*****
60	$\text{aC}-\text{CH}_n-\text{OOC}-\text{H}$ (n in 1..2)	Benzyl formate (1)	*****	0.2238	1.7860	0.004195	-3.40	3.020	4.145	*****	*****
61	$\text{aC}-\text{CH}_m-\text{NO}_2$ (n in 1..2)	Phenyl nitromethane (1)	*****	0.5390	*****	*****	*****	*****	*****	*****	*****
62	$\text{aC}-\text{CH}_n-\text{CONH}_2$ (n in 1..2)	Phenyl ethanamide (1)	2.2421	-0.2197	*****	*****	*****	*****	*****	*****	*****
63	$\text{aC}-\text{CH}_n-\text{OOC}$ (n in 1..2)	Benzyl acetate (1)	-0.6997	0.0886	1.1629	-0.000384	-7.02	1.556	4.066	*****	*****
64	$\text{aC}-\text{CH}_n-\text{COO}$ (n in 1..2)	Methyl phenyl acetate (1)	-0.2636	0.0352	*****	*****	*****	*****	*****	*****	*****
65	$\text{aC}-\text{SO}_2-\text{OH}$	Benzenesulfonic acid (1)	-1.1057	*****	*****	*****	*****	*****	*****	*****	*****
66	$\text{aC}-\text{CH}(\text{CH}_3)_2$	Cumene (1)	0.0642	0.0196	0.1565	-0.001446	-2.04	1.238	-0.751	1.030	-0.270
67	$\text{aC}-\text{C}(\text{CH}_3)_3$	<i>tert</i> -Butylbenzene (1)	0.0790	0.0494	0.8016	-0.006495	-5.70	0.354	-0.192	*****	-0.878
68	$\text{aC}-\text{CF}_3$	Perfluorotoluene (1)	-10.8058	-1.5974	*****	*****	*****	*****	*****	*****	*****
69	$(\text{CH}_n=\text{C})(\text{cyclic})-\text{CHO}$ (n in 0..2)	Furfural (1)	-1.0516	0.4267	2.4070	-0.002650	0.39	-6.438	-12.517	*****	-1.670
70	$(\text{CH}_n=\text{C})_{\text{cyc}}-\text{COO}-\text{CH}_m$ (n, m in 0..3)	Methyl furanurate (1)	-6.9427	0.0879	*****	*****	*****	*****	*****	*****	*****
71	$(\text{CH}_n=\text{C})_{\text{cyc}}-\text{CO}-$ (n in 0..2)	2-Acetyl furan (1)	0.6572	0.6115	*****	*****	*****	*****	*****	*****	*****
72	$(\text{CH}_n=\text{C})_{\text{cyc}}-\text{CH}_3$ (n in 0..2)	1,2-Dimethylcyclopentene (2)	0.0416	0.0173	-0.2509	-0.000624	0.03	28.972	24.560	*****	2.235
73	$(\text{CH}_n=\text{C})_{\text{cyc}}-\text{CH}_2$ (n in 0..2)	2-Ethylfuran (1)	-0.3151	-0.0504	-1.1019	0.003921	-4.43	-22.533	-12.044	*****	0.961
74	$(\text{CH}_n=\text{C})_{\text{cyc}}-\text{CN}$ (n in 0..2)	3-Cyanofuran (1)	1.5819	-0.2474	*****	*****	*****	*****	*****	*****	*****
75	$(\text{CH}_n=\text{C})_{\text{cyc}}-\text{Cl}$ (n in 0..2)	2-Chlorofuran (1)	-0.8604	-0.5736	*****	*****	*****	*****	*****	*****	*****
76	$\text{CH}_{\text{cyc}}-\text{CH}_3$	Methylcyclopentane (1)	-0.1326	-0.1210	-0.1233	0.000779	2.79	4.178	4.452	0.096	0.033
77	$\text{CH}_{\text{cyc}}-\text{CH}_2$	Ethylcyclohexane (1)	-0.4669	-0.0148	0.3816	0.001694	-2.95	5.332	4.428	-0.428	-1.137
78	$\text{Ch}_{\text{cyc}}-\text{CH}$	Isopropylcyclopentane (1)	-0.3548	0.1395	0.1093	0.000124	6.19	6.084	-4.128	0.153	2.421
79	$\text{Ch}_{\text{cyc}}-\text{C}$	<i>tert</i> -Butylcyclohexane (1)	-0.1727	0.1829	*****	*****	*****	*****	*****	*****	*****
80	$\text{Ch}_{\text{cyc}}-\text{CH}=\text{CH}_n$ (n in 1..2)	Vinylcyclopentane (1)	0.6817	-0.1192	*****	*****	*****	*****	*****	*****	*****
81	$\text{Ch}_{\text{cyc}}-\text{C}=\text{CH}_n$ (n in 1..2)	Limonene (1)	-1.0631	-0.0455	-0.2832	0.002114	-16.97	6.768	10.390	*****	*****
82	$\text{Ch}_{\text{cyc}}-\text{Cl}$	Chloro cyclopentane (1)	0.5124	0.2667	*****	*****	*****	*****	*****	*****	*****
83	$\text{Ch}_{\text{cyc}}-\text{F}$	Fluoro cyclohexane (1)	2.8497	-0.1899	*****	*****	*****	*****	*****	*****	*****
84	$\text{Ch}_{\text{cyc}}-\text{OH}$	Cyclohexanol (1)	1.3691	-0.3179	0.8973	0.004640	-7.73	-3.024	-8.050	2.134	*****
85	$\text{Ch}_{\text{cyc}}-\text{NH}_2$	Cyclohexylamine (1)	1.5069	-0.3576	-0.9610	0.000039	-2.50	2.046	3.446	-4.607	0.328
86	$\text{Ch}_{\text{cyc}}-\text{NH}-\text{CH}_n$ (n in 0..3)	<i>N</i> -methylcyclohexylamine (1)	0.0370	-0.7458	-2.0833	-0.014535	-51.50	-11.965	14.531	*****	0.402
87	$\text{Ch}_{\text{cyc}}-\text{N}-\text{CH}_n$ (n in 0..3)	<i>N,N</i> -dimethylcyclohexanamine (1)	*****	0.1218	*****	*****	*****	*****	*****	*****	*****
88	$\text{Ch}_{\text{cyc}}-\text{SH}$	Cyclohexanethiol (1)	-0.3312	-0.0569	-0.6447	-0.000199	-2.00	-16.723	-7.569	*****	-0.878
89	$\text{Ch}_{\text{cyc}}-\text{CN}$	Cyanocyclopentane (1)	*****	0.4649	*****	*****	*****	*****	*****	*****	*****
90	$\text{Ch}_{\text{cyc}}-\text{COOH}$	Cyclopropanecarboxylic acid (1)	-2.0822	0.1506	*****	*****	*****	*****	*****	*****	*****
91	$\text{Ch}_{\text{cyc}}-\text{CO}$	Methyl cyclohexyl ketone (1)	0.7743	0.1300	*****	*****	*****	*****	*****	-0.616	*****
92	$\text{Ch}_{\text{cyc}}-\text{NO}_2$	Nitrocyclohexane (1)	-0.8578	0.6540	*****	*****	*****	*****	*****	*****	*****
93	$\text{Ch}_{\text{cyc}}-\text{S}-$	Methyl cyclopentyl sulfide (1)	-0.8638	0.0043	*****	*****	*****	*****	*****	*****	*****
94	$\text{Ch}_{\text{cyc}}-\text{CHO}$	Cyclohexanecarboxaldehyde (1)	0.5076	-0.2692	*****	*****	*****	*****	*****	*****	*****
95	$\text{Ch}_{\text{cyc}}-\text{O}-$	Methoxycyclohexane (1)	-0.3978	-0.2787	*****	*****	*****	*****	*****	*****	*****

	Group	Example	T_{m2j}	T_{b2j}	T_{c2j}	P_{c2j}	V_{c2j}	G_{l2j}	H_{l2j}	H_{v2j}	H_{lu2j}
96	Ch _{cyC} -OOCH	Cyclohexyl ester formic acid (1)	*****	-0.2107	*****	*****	*****	*****	*****	*****	*****
97	Ch _{cyC} -COO	Ethyl cyclobutyrate (1)	*****	0.0926	*****	*****	*****	*****	*****	*****	*****
98	Ch _{cyC} -OOC	Cyclohexyl acetate (1)	-0.4666	-0.4495	-0.3450	-0.000692	-12.03	4.358	-15.751	*****	*****
99	C _{cyC} -CH ₃	1,1-Dimethyl-cyclohexane (2)	0.1737	0.0722	0.1607	0.001235	1.95	0.107	0.238	0.808	-1.237
100	C _{cyC} -CH ₂	1-Ethyl-1-methyl-cyclopentane (1)	-1.9233	0.0319	0.1090	-0.000610	-5.17	18.755	21.498	0.585	*****
101	C _{cyC} -OH	1-Methylcyclopentanol (1)	0.7334	-0.6775	-2.1303	-0.004683	-14.40	-18.970	-21.975	*****	0.235
102	>N _{cyC} -CH ₃	N-methyl-2-pyrrolidone (1)	-0.0383	0.0604	-0.0003	0.000058	*****	*****	*****	*****	*****
103	>N _{cyC} -CH ₂	N-ethylpyrrole (1)	1.0497	-0.3080	*****	*****	*****	*****	*****	*****	*****
104	AROMRINGS ¹ s ²	2-Methyl-phenol (1), 2-Et-toluene (1)	-0.6388	-0.1590	-0.3161	0.000522	2.86	1.577	1.486	1.164	-1.470
105	AROMRINGS ¹ s ³	3-Methyl-phenol (1), 3-Et-toluene (1)	-0.6218	0.0217	-0.0693	0.001790	6.54	-1.037	0.294	-1.910	-1.059
106	AROMRINGS ¹ s ⁴	4-Methyl-phenol (1), 4-Et-toluene (1)	0.9840	0.1007	0.0803	0.000467	3.70	-0.709	0.384	0.331	1.244
107	AROMRINGS ¹ s ² s ³	1,2,3-Trimethylbenzene (1)	-0.2762	-0.1647	1.0088	-0.005598	-9.58	7.731	5.743	1.433	0.473
108	AROMRINGS ¹ s ² s ⁴	1,2,4-Trihydroxybenzene (1)	-0.3689	-0.1387	0.0908	0.000255	-2.05	-2.767	-0.449	0.313	-0.302
109	AROMRINGS ¹ s ³ s ⁵	3,5-Diethyltoluene (1)	-0.3841	-0.1314	-0.6412	0.004090	-7.67	-2.148	-7.538	-0.117	-2.530
110	AROMRINGS ¹ s ² s ³ s ⁴	3-Ethyl-1,2,4-trimethylbenzene (1)	1.7722	0.2745	2.1116	-0.007612	-7.04	14.226	12.710	*****	-1.736
111	AROMRINGS ¹ s ² s ³ s ⁵	1,2,3,5-Tetramethylbenzene (1)	0.4553	0.1645	0.9353	-0.001811	-0.04	4.926	5.220	*****	-2.246
112	AROMRINGS ¹ s ² s ⁴ s ⁵	1,2,4,5-Tetramethylbenzene (1)	2.0561	0.0754	0.6241	-0.000500	-0.04	-0.474	-1.340	*****	8.034
113	PYRIDINES ²	2-Methylpyridine (1)	-0.5769	-0.1196	-1.0256	0.007006	8.68	-9.713	-9.644	-1.683	-0.786
114	PYRIDINES ³	3-Methylpyridine (1)	-0.2556	0.0494	0.5784	0.007006	8.68	-2.523	-2.446	0.277	3.671
115	PYRIDINES ⁴	4-Methylpyridine (1)	1.6282	0.1344	0.6595	0.001283	14.28	-4.703	-6.466	0.397	5.975
116	PYRIDINES ² s ³	2,3-Dimethylpyridine (1)	-0.1341	0.0032	*****	*****	*****	*****	*****	-0.939	*****
117	PYRIDINES ² s ⁴	2,4-Dimethylpyridine (1)	-1.6848	-0.0817	*****	*****	*****	*****	*****	-1.269	*****
118	PYRIDINES ² s ⁵	2,5-Dimethylpyridine (1)	-0.9802	-0.1564	*****	*****	*****	*****	*****	-1.719	*****
119	PYRIDINES ² s ⁶	2,6-Dimethylpyridine (1)	0.3018	-0.5176	-2.2773	0.008029	-50.26	-16.570	-17.778	-3.419	-1.487
120	PYRIDINES ³ s ⁴	3,4-Dimethylpyridine (1)	0.1018	0.5477	*****	*****	*****	*****	*****	1.742	*****
121	PYRIDINES ³ s ⁵	3,5-Dimethylpyridine (1)	0.2811	0.3533	*****	*****	*****	*****	*****	0.572	*****
122	PYRIDINES ² s ³ s ⁶	2,3,6-Trimethylpyridine (1)	-0.3189	-0.3888	*****	*****	*****	*****	*****	-2.744	*****

Table A.3: Third Order Group Contribution Property Data (Marrero & Gani, 2001)

Third-order groups and their contributions along with sample assignments

Group	Example	T_{m3k}	T_{h3k}	T_{c3k}	P_{c3k}	V_{c3k}	G_{E3k}	H_{E3k}	H_{Tm3k}	
1	HOOC-(CH ₂) _m -COOH ($m > 2, n$ in 0..2)	1,5-Pentanedioic acid (1)	-1.5257	1.6498	-1.6986	0.001544	-3.72	-4.708	-6.572	-7.583
2	NH ₂ -(CH ₂) _m -COOH ($m > 2, n$ in 0..2)	4-Aminobutyric acid (1)	11.2271	*****	*****	*****	*****	*****	*****	*****
3	NH ₂ -(CH ₂) _m -OH ($m > 2, n$ in 0..2)	4-Aminobutanol (1)	0.7732	1.0750	0.4950	0.000728	-23.74	3.079	4.171	-4.840
4	OH-(CH ₂) _m -OH ($m > 2, n$ in 0..2)	1,9-Nonanediol (1)	0.6674	0.7193	0.1725	-0.000327	-0.84	7.536	5.411	-0.272
5	OH-(CH ₂) _k -O-(CH ₂) _m -OH ($m, k > 0; p, n$ in 0..2)	Dipropylene glycol (1)	-0.1073	1.1867	6.6872	0.001937	1.44	-8.397	-8.651	1.661
6	OH-(CH ₂) _k -S-(CH ₂) _m -OH ($m, k > 0; p, n$ in 0..2)	2,2'-Diethyl-dihydroxy sulfide (1)	-1.3891	*****	2.6769	0.003792	-1.62	10.194	8.164	-3.479
7	OH-(CH ₂) _k -NHX-(CH ₂) _m -OH ($m, k > 0; p, n, x$ in 0..2)	Diethanolamine (1)	-0.0781	0.2991	*****	0.003254	-0.69	1.662	1.753	0.301
8	CH ₃ -O-(CH ₂) _m -OH ($m > 2; n, p$ in 0..2)	Butoxypropanol (1)	*****	-0.4605	*****	*****	*****	*****	*****	*****
9	NH ₂ -(CH ₂) _m -NH ₂ ($m > 2; n$ in 0..2)	1,5-Diaminopentane (1)	-0.0604	0.0060	-4.3195	0.006734	6.69	4.100	0.371	5.666
10	NH _k -(CH ₂) _m -NH ₂ ($m > 2; k$ in 0..1; n in 0..2)	<i>N,N</i> -dimethylpropylenediamine (1)	-1.1888	-0.1819	*****	*****	*****	*****	*****	*****
11	SH-(CH ₂) _m -SH ($m > 2; n$ in 0..2)	1,5-Pentanedithiol (1)	0.6669	0.4516	*****	*****	*****	*****	*****	*****
12	NC-(CH ₂) _m -CN ($m > 2$)	Glutaronitrile (1)	-0.3798	1.3440	0.0834	-0.011090	-36.89	-7.035	7.782	-0.607
13	COO-(CH ₂) _m -OOC ($m > 2; n$ in 0..2)	Glyceryl tridodecanoate (1)	-2.6542	*****	*****	*****	*****	*****	*****	*****
14	aC-(CH _n =CH _m) _{cyc} (fused rings) (n, m in 0..1)	Indene (1), Acenaphylene (2)	0.2479	-0.3741	-0.0185	0.000851	-8.87	-1.601	2.689	-2.703
15	aC-aC (different rings)	Biphenylene (2), Biphenyl (1)	1.1395	-0.4961	6.1894	-0.040100	-26.26	-4.459	-4.558	-0.385
16	aC-CH _{n,cyc} (different rings) (n in 0..1)	Cyclohexylbenzene (1)	0.0570	-0.4574	-0.2474	-0.005826	-8.55	-5.267	-5.914	-0.442
17	aC-CH _{n,cyc} (fused rings) (n in 0..1)	Tetralin (2), Indane (2)	-0.5640	-0.1736	0.5060	-0.003746	-11.56	-4.203	-4.863	-0.143
18	aC-(CH _n) _m -aC (different rings) ($m > 1; n$ in 0..2)	Bibenzyl (1)	1.9902	0.3138	3.0321	0.003007	9.73	1.318	0.084	5.377
19	aC-(CH _n) _m -CH _{2,cyc} (different rings) ($m > 0; n$ in 0..2)	1-Cyclopentyl-3-phenylpropane (1)	*****	0.5928	*****	*****	*****	*****	*****	*****
20	CH _{cyc} -CH _{cyc} (different rings)	Cyclohexylcyclohexane (1)	0.5460	0.4387	2.1761	0.002745	7.72	-67.517	-66.870	*****
21	CH _{cyc} -(CH ₂) _m -CH _{cyc} (different rings) ($m > 0; n$ in 0..2)	1,2-Dicyclohexylethane (1)	0.4497	0.5632	*****	*****	*****	*****	*****	*****
22	CH multiring	Hexahydroindan (2), Decalin (2)	0.6647	0.1415	0.4963	-0.000985	-3.33	*****	*****	0.223
23	C multiring	Spiropentane (1)	0.0792	*****	*****	*****	*****	*****	*****	*****
24	aC-CH _m -aC (different rings) (m in 0..2)	Diphenylmethane (1)	0.6457	0.2391	0.1174	-0.002673	-4.67	-0.729	0.866	-0.958
25	aC-(CH _m =CH _n) _{cyc} -aC (different rings) (m, n in 0..2)	1,2-Diphenylethylene (1)	0.9608	0.7192	0.7039	-0.004661	14.31	-0.702	-2.291	3.275
26	(CH _w =C) _{cyc} -CH=CH-(C=CH _n) _{cyc} (different rings)	1,2-Furanyl ethene (1)	16.2235	*****	*****	*****	*****	*****	*****	*****
27	(CH _w =C) _{cyc} -CH _p -(C=CH _n) _{cyc} (different rings)	Difuranyl methane (1)	16.8558	*****	*****	*****	*****	*****	*****	*****
28	aC-CO-aC (different rings)	Benzophenone (1)	-1.0394	1.0171	-0.2678	-0.001837	-7.05	11.125	7.108	-4.091
29	aC-CH _m -CO-aC (different rings) (m in 0..2)	Benzyl phenone (1)	-0.4486	0.9674	*****	*****	*****	*****	*****	*****
30	aC-CO-(C=CH _n) _{cyc} (different rings) (n in 0..1)	Phenyl-2-furanyl-methanone (1)	-0.1376	0.1126	*****	*****	*****	*****	*****	*****
31	aC-CO-CO-aC (different rings)	Diphenylethanedione (1)	0.4361	0.9317	*****	*****	*****	*****	*****	-3.687
32	aC-CO _{cyc} (fused rings)	Phenolphthalein (1)	3.6847	0.5031	*****	*****	*****	*****	*****	2.047
33	aC-CO-(CH _n) _m -CO-aC (different rings) ($m > 0; n$ in 0..2)	1,4-Diphenyl-1,4-butanedione (1)	4.9038	*****	*****	*****	*****	*****	*****	7.327
34	aC-CO-CH _{n,cyc} (different rings) (n in 0..1)	Cyclohexyl phenyl methanone (1)	-7.0038	*****	*****	*****	*****	*****	*****	*****
35	aC-CO-NH _n -aC (different rings) (n in 0..1)	<i>N</i> -phenyl benzamide (1)	5.9653	*****	*****	*****	*****	*****	*****	2.510
36	aC-NH _n CONH _m -aC (different rings) (n, m in 0..1)	<i>N,N'</i> -diphenylurea (1)	1.5629	*****	*****	*****	*****	*****	*****	0.018
37	aC-CO-N _{cyc} (different rings)	<i>N</i> -phenonyl-piperidine (1)	-9.1856	*****	*****	*****	*****	*****	*****	*****
38	aC-S _{cyc} (fused rings)	Dibenzothiophene (2)	0.2612	0.2242	3.5541	0.004600	12.60	8.333	9.212	-0.784
39	aC-S-aC (different rings)	Diphenyl sulfide (1)	-1.8403	0.0185	*****	*****	*****	*****	*****	*****
40	aC-PO _n -aC (different rings) (n in 0..4)	Triphenylphosphate (3)	0.0393	*****	*****	*****	*****	*****	*****	*****
41	aC-SO _n -aC (different rings) (n in 1..4)	Diphenyl sulfone (1)	0.9514	-0.0850	*****	*****	*****	*****	*****	-2.485
42	aC-NH _{n,cyc} (fused rings) (n in 0..1)	Carbazole (2)	3.4983	1.1457	3.5541	0.017201	0.44	-2.221	-16.080	0.196
43	aC-NH-aC (different rings)	Diphenylamine (1)	-0.3048	0.5768	0.9519	0.008484	1.42	-0.596	-1.994	1.934
44	aC-(C=N) _{cyc} (different rings)	Phenyl-3-pyrazole (1)	-1.3060	-0.5335	*****	*****	*****	*****	*****	*****
45	aC-(N=CH _n) _{cyc} (fused rings) (n in 0..1)	Benzoxazole (1)	-4.9289	-5.2736	*****	*****	*****	*****	*****	-0.599
46	aC-(CH _n =N) _{cyc} (fused rings) (n in 0..1)	Benzoisoxazole (1)	-10.1007	*****	*****	*****	*****	*****	*****	*****

Group	Example	T_{m3k}	T_{h3k}	T_{c3k}	P_{c3k}	V_{c3k}	G_{f3k}	H_{f3k}	H_{fm3k}	
47	aC-O-CH _n -aC (different rings) (<i>n</i> in 0..2)	Benzyl phenyl ether (1)	1.0834	0.6571	*****	*****	*****	*****	*****	
48	aC-O-aC (different rings)	Diphenyl ether (1)	-0.4803	-0.8252	-0.9785	0.001162	-2.63	2.668	-5.074	1.193
49	aC-CH _n -O-CH _m -aC (different rings) (<i>n, m</i> in 0..2)	Benzyl ether (1)	-3.2676	0.2790	-1.4002	-0.004716	28.42	-4.229	-2.303	-3.971
50	aC-O _{ovc} (fused rings)	Benzoxazole (1)	-0.3545	-0.6848	*****	*****	*****	*****	*****	-1.153
51	AROMFUSED[2]	Naphthalene (2)	0.2825	0.0441	-1.0095	-0.001332	-6.88	1.993	1.904	0.694
52	AROMFUSED[2]s ¹	1-Methylnaphtalene (1)	-1.2836	-0.1666	0.1605	-0.002030	-3.17	-2.940	-2.274	-3.699
53	AROMFUSED[2]s ²	2,7-Dimethylnaphtalene (2)	0.3378	-0.2692	-0.6765	-0.002436	-3.85	-1.873	-1.316	2.037
54	AROMFUSED[2]s ² s ³	2,3-Dimethylnaphtalene (1)	1.8941	-0.2807	*****	*****	*****	*****	*****	2.150
55	AROMFUSED[2]s ¹ s ⁴	1,4-Dimethylnaphtalene (1)	-2.7585	-0.3294	*****	*****	*****	*****	*****	*****
56	AROMFUSED[2]s ¹ s ²	1,2-Dimethylnaphtalene (1)	-3.0362	-0.2931	*****	*****	*****	*****	*****	*****
57	AROMFUSED[2]s ¹ s ³	1,3-Dimethylnaphtalene (1)	-3.2228	-0.3360	*****	*****	*****	*****	*****	*****
58	AROMFUSED[3]	Phenylene (3), Pyrene (2)	1.6600	0.0402	-1.0430	0.004695	35.21	3.896	5.819	1.176
59	AROMFUSED[4a]	Anthracene (1)	7.0402	1.0466	3.3011	0.015244	-6.96	13.843	11.387	5.027
60	AROMFUSED[4a]s ¹	9-Methylanthracene (1)	-3.3463	-7.8521	*****	*****	*****	*****	*****	*****
61	AROMFUSED[4a]s ¹ s ⁴	9,10-Dimethylanthracene (1)	6.8373	*****	*****	*****	*****	*****	*****	*****
62	AROMFUSED[4p]	Phenanthrene (1), Pyrene (2)	-1.5856	0.9126	2.8885	0.007280	-24.02	-16.040	-19.089	-3.417
63	AROMFUSED[4p]s ¹ s ⁴	9,10-Dimethylphenanthrene (1)	2.0821	*****	*****	*****	*****	*****	*****	*****
64	PYRIDINE.FUSED[2]	Quinoline (1)	-4.4725	-0.9432	1.1251	-0.005369	63.29	8.688	13.586	-4.967
65	PYRIDINE.FUSED[2-iso]	Isoquinoline (1)	-2.5898	-0.5844	3.9241	-0.011207	-2.71	-5.112	-0.314	-2.587
66	PYRIDINE.FUSED[4]	Acridine (1)	1.0358	0.1733	7.7134	-0.001275	-12.04	20.073	15.786	-1.365

Table A. 4: First Order GC Data for Acentric Factors and Liquid Molar Volume (Constantinou et al., 1995)

First-order groups and their contributions for estimation of acentric factors and liquid molar volume		
Group	w_{1i}	v_{1i} ($\text{m}^3 \text{ kmol}^{-1}$)
CH ₃	0.29602	0.02614
CH ₂	0.14691	0.01641
CH	-0.07063	0.00711
C	-0.35125	-0.00380
CH ₂ =CH	0.40842	0.03727
CH=C	0.25224	0.02692
CH ₂ =C	0.22309	0.02697
CH=C	0.23492	0.01610
C-C	-0.21017	0.00296
CH ₂ =C=CH	0.73865	0.04340
ACH	0.15188	0.01317
AC	0.02725	0.00440
ACCH ₁	0.33409	0.02888
ACCH ₂	0.14598	0.01916
ACCH	-0.08807	0.00993
OH	1.52370	0.00551
ACOH	0.73657	0.01133
CH ₂ CO	1.01522	0.03655
CH ₂ CO	0.63264	0.02816
CHO	0.96265	0.02002
CH ₂ COO	1.13257	0.04500
CH ₂ COO	0.75574	0.03567
HCOO	0.76454	0.02667
CH ₂ O	0.52646	0.03274
CH ₂ O	0.44184	0.02311
CH-O	0.21808	0.01799
FCH ₂ O	0.50922	0.02059
CH ₂ NH ₂	0.79963	0.02646
CHNH ₂	****	0.01952
CH ₂ NH	0.95344	0.02674
CH ₂ NH	0.55018	0.02318
CHNH	0.38623	0.01813
CH ₂ N	0.38447	0.01913
CH ₂ N	0.07508	0.01683
ACNH ₂	0.79337	0.01365
C ₂ H ₄ N	****	0.06082
C ₂ H ₃ N	****	0.05238
CH ₂ CN	****	0.03313
COOH	1.67037	0.02232
CH ₂ Cl	0.57021	0.03371
CHCl	****	0.02663
CCl	****	0.02020
CHCl ₂	0.71592	0.04682
CCl ₂	****	****
CCl ₃	0.61662	0.06202
ACCl	****	0.02414

Group	w_{1i}	v_{1i} ($\text{m}^3 \text{ kmol}^{-1}$)
CH ₂ NO ₂	****	0.03375
CHNO ₂	****	0.02620
ACNO ₂	****	0.02505
CH ₂ SH	****	0.03446
I	0.23323	0.02791
Br	0.27778	0.02143
CH=C	0.61802	****
C=C	****	0.01451
Cl-(C-C)	****	0.01533
ACF	0.26254	0.01727
HCON(CH ₂) ₂	****	****
CF ₃	0.50023	****
CF ₂	****	****
CF	****	****
COO	****	0.01917
CCl ₂ F	0.50260	0.05384
HCClF	****	****
CClF ₂	0.54685	0.05383
F (except as above)	0.43796	****
CONH ₂	****	****
CONHCH ₃	****	****
CONHCH ₂	****	****
CON(CH ₃) ₂	****	0.05477
CONCH ₃ CH ₂	****	****
CON(CH ₂) ₂	****	****
C ₂ H ₃ O ₂	****	0.04104
C ₂ H ₄ O ₂	****	****
CH ₃ S	****	0.03484
CH ₂ S	0.42753	0.02732
CHS	****	****
C ₆ H ₅ S	****	****
C ₄ H ₅ S	****	****

Table A. 5: Second Order GC Data for Acentric Factors and Liquid Molar Volume (Constantinou et al., 1995)

Second-order groups and their contributions for estimation of acentric factors and liquid molar volumes at 298 K			
Groups	w_{2j}	v_{2j} ($\text{m}^3 \text{ kmol}^{-1}$)	Sample assignments (occurrences)
$(\text{CH}_3)_2\text{CH}$	0.01740	0.00133	2-Methylpentane (1)
$(\text{CH}_3)_3\text{C}$	0.01922	0.00179	2,2-Dimethylpentane (1), 2,2,4,4-tetramethylpentane (2)
$\text{CH}(\text{CH}_3)\text{CH}(\text{CH}_3)$	-0.00475	-0.00203	2,3-Dimethylpentane (1), 2,3,4-tetramethylpentane (2)
$\text{CH}(\text{CH}_3)\text{C}(\text{CH}_3)_2$	-0.02883	-0.00243	2,2,3-Dimethylpentane (1), 2,2,3,4,4-pentamethylpentane (2)
$\text{C}(\text{CH}_3)_2\text{C}(\text{CH}_3)_2$	-0.08632	-0.00744	2,2,3,3-Dimethylpentane (1), 2,2,3,3,4,4-hexamethylpentane (2)
3-Membered ring ^a	0.17563	****	Cyclopropane (1)
4-Membered ring ^a	0.22216	****	Cyclobutane (1)
5-Membered ring ^a	0.16284	0.00213	Cyclopentane (1), ethylcyclopentane (1)
6-Membered ring ^a	-0.03065	0.00063	Cyclohexane (1), methylcyclohexane (1)
7-Membered ring ^a	-0.02094	-0.00519	Cycloheptane (1), ethylcycloheptane (1)
$\text{CH}_n-\text{CH}_m-\text{C}_p=\text{C}_k$ $k,n,m,p \in (0, 2)$	0.01648	-0.00188	1,3-Butadiene (1)
$\text{CH}_3-\text{CH}_m=\text{CH}_n$ $m,n \in (0, 2)$	0.00619	0.00009	2-Butene (2), 2-methyl-2-butene (3)
$\text{CH}_2-\text{CH}_m=\text{CH}_n$ $m,n \in (0, 2)$	-0.0115	0.00012	1,4-Pentadiene (2)
$\text{CH}-\text{CH}_m=\text{CH}_n$ or $\text{C}-\text{CH}_m=\text{CH}_n$ $m,n \in (0, 2)$	0.02778	0.00142	4-Methyl-2-pentene (1)
Alicyclic side chain $\text{C}_{\text{cyclic}}\text{C}_m$ $m > 1$	-0.11024	-0.00107	Ethylcyclohexane (1), propylcycloheptane (1)
CH_3CH_3	-0.1124	****	Ethane (only)
CHCHO or CCHO	****	-0.00009	2-Methyl butanaldehyde (1)
CH_3COCH_2	-0.20789	-0.00030	2-Pentanone (1)
CH_3COCH or CH_3COC	-0.16571	-0.00108	3-Methyl-2-pentanone (1)
$\text{C}_{\text{cyclic}}=\text{O}$	****	-0.00111	Cyclohexanone (1)
ACCHO	****	-0.00036	Benzaldehyde (1)
CHCOOH or CCOOH	0.08774	-0.00050	2-Methyl butanoic acid (1)
ACCOOH	****	0.00777	Benzoic acid (1)
CH_3COOCH or CH_3COOC	-0.26623	0.00083	Isopropyl acetate (1)
COCH_2COO , COCHCOO or COCCOO	****	0.00036	Ethyl acetoacetate (1)
$\text{CO}-\text{O}-\text{CO}$	0.91939	0.00198	Propanoic anhydrite (1)
ACCOO	****	0.00001	Benzoic acid ethyl ester (1)
CHOH	0.03654	-0.00092	2-Butanol (1)
COH	0.21106	0.00175	2-Methyl-2-butanol (1)
$\text{CH}_m(\text{OH})\text{C}_n(\text{OH})$ $m,n \in (0, 2)$	****	0.00235	1,2,3-Propanetriol (2)
$\text{CH}_m\text{cyclic}-\text{OH}$ $m \in (0, 1)$	****	-0.00250	Cyclopentanol (1)
$\text{CH}_m(\text{OH})\text{CH}_n(\text{NH}_p)$ $m,n,p \in (0, 2)$	****	0.00046	1-Amino-2-butanol (1), 1-hydroxy-N-methylbutylamine (1)

Groups	w_{2j}	v_{2j} ($\text{m}^3 \text{ kmol}^{-1}$)	Sample assignments (occurrences)
$\text{CH}_m(\text{NH}_2)\text{CH}_n(\text{NH}_2)$ $m, n \in (0, 2)$	****	****	1,2-Propanodiamine (1)
$\text{CH}_{m \text{ cyclic}}-\text{NH}_p-\text{CH}_n \text{ cyclic}$ $m, n, p \in (0, 2)$	-0.13106	-0.00179	Pyrrolidine (1)
$\text{CH}_m-\text{O}-\text{CH}_n=\text{CH}_p$ $m, n, p \in (0, 2)$	****	-0.00206	Ethyl vinyl ether (1)
$\text{AC}-\text{O}-\text{CH}_m$ $m(0, 3)$	****	0.01203	Ethyl phenyl ether (1)
$\text{CH}_{m \text{ cyclic}}-\text{S}-\text{CH}_n \text{ cyclic}$ $m, n \in (0, 2)$	-0.01509	-0.00023	Tetrahydrothiophene (1)
$\text{CH}_m=\text{CH}_n-\text{F}$ $m, n \in (0, 2)$	****	****	1-Fluoro-1-propene (1)
$\text{CH}_m=\text{CH}_n-\text{Br}$ $m, n \in (0, 2)$	****	-0.0058	1-Bromo-1-propene (1)
$\text{CH}_m=\text{CH}_n-\text{I}$ $m, n \in (0, 2)$	****	****	1-Iodo-1-propene (1)
ACBr	-0.03078	0.00178	Bromotoluene (1)
ACI	0.00001	0.00171	Iodotoluene (1)
$\text{CH}_m(\text{NH}_2)-\text{COOH}$ $m \in (0, 2)$	****	****	2-Aminohexanoic acid (1)

Table A. 6: Regressed Parameters for Different Atom Types in the CI Method (Gani et al., 2005)

parameter type	properties							
	T_m (10^{-1})	T_b (10^{-1})	T_c (10^{-2})	P_c (10^{-3})	V_c	H_f	G_f	H_{fus} (10^{-1})
$a(H)$	-1.951 16	-1.194 61	-44.252 84	2.022 97	7.119 75	-34.777 51	-15.256 65	-0.141 97
$a(Cl)$	17.742 44	14.001 77	448.531 72	5.370 86	43.403 16	-66.442 25	-47.252 86	3.800 52
$a(Br)$	44.965 78	24.031 95	739.233 87	-6.189 42	51.377 39	-40.041 62	-30.932 67	5.007 44
$a(F)$	-8.182 62	-0.791 56	-9.098 94	7.725 04	20.807 61	-238.125 24	-222.598 77	0.521 16
$a(I)$	43.472 78	35.274 55	312.383 03	-10.784 26	68.446 31	10.071 69	11.310 87	5.686 94
$a(N)$	28.882 43	16.237 96	584.222 14	4.093 67	39.903 16	92.740 20	67.399 56	4.409 31
$a(O)$	19.879 42	9.283 53	372.271 66	-1.389 01	18.047 65	-176.070 06	-168.720 71	3.621 08
$a(P)$	3.304 41	2.481 35	795.371 60	N/A	-82.464 85	-243.529 50	N/A	-11.318 61
$a(S)$	26.658 39	17.769 81	780.776 34	-8.430 30	32.232 63	9.576 91	-1.173 46	4.600 86
$a(C)$	10.864 15	11.312 90	324.582 68	5.499 90	31.797 84	40.155 90	35.307 60	2.214 78
$a(Si)$	-1.340 33	3.731 42	N/A	N/A	N/A	N/A	N/A	N/A
b	2.631 05	-9.382 97	-327.749 36	2.129 80	3.096 75	-7.395 81	-21.519 51	-3.663 88
c	-10.868 99	4.604 18	125.059 14	6.521 88	7.874 95	11.717 23	15.348 11	3.113 80
d	0.000 00	18.371 91	388.551 35	-18.972 22	8.673 18	61.926 11	97.288 21	1.099 72
E	1474.5	2225.4	23123.9	5982.7	7.95	5.549	-34.967	-28.06
G				108.998				