Development of a Group Contribution Method for the Prediction of Normal Boiling Points of Non-electrolyte Organic Compounds

By

Yash Nannoolal

[B.Sc. (Eng.)]

University of Kwazulu-Natal Durban

For the degree Master of Science (Chemical Engineering)

January 2004

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Abstract

Physical properties are fundamental to all chemical, biochemical and environmental industries. One of these properties is the normal boiling point of a compound. However, experimental values in literature are quite limited and measurements are expensive and time consuming. For this reason, group contribution estimation methods are generally used. Group contribution is the simplest form of estimation requiring only the molecular structure as input. Consequently, the aim of this project was the development of a reliable group contribution method for the estimation of normal boiling points of non-electrolytes applicable for a broad range of components.

A literature review of the available methods for the prediction of the normal boiling points from molecular structure only, was initially undertaken. From the review, the Cordes and Rarey (2002) method suggested the best scientific approach to group contribution. This involved defining the structural first-order groups according to its neighbouring atoms. This definition also provided knowledge of the neighbourhood and the electronic structure of the group. The method also yielded the lowest average absolute deviation and probability of prediction failure. Consequently, the proposed group contribution method was then developed using the Cordes and Rarey method as a starting point. The data set included experimental data for approximately 3000 components, 2700 of which were stored in the Dortmund Data Bank (DDB) and about 300 stored in Beilstein.

The mathematical formalism was modified to allow for separate examination and regression of individual contributions using a meta-language filter program developed specifically for this purpose. The results of this separate examination lead to the detection of unreliable data, the re-classification of structural groups, and introduction of new structural groups to extend the range of the method.

The method was extended using steric parameters, additional corrections and group interaction parameters. Steric parameters contain information about the greater neighbourhood of a carbon. The additional corrections were introduced to account for certain electronic and structural effects that the first-order groups could not capture.

Group interactions were introduced to allow for the estimation of complex multifunctional compounds, for which previous methods gave extraordinary large deviations from experimental findings. Several approaches to find an improved linearization function did not lead to an improvement of the Cordes and Rarey method.

The results of the new method are extensively compared to the work of Cordes and Rarey and currently-used methods and are shown to be far more accurate and reliable. Overall, the proposed method yielded an average absolute deviation of 6.50K (1.52%) for a set of 2820 components. For the available methods, Joback and Reid produced an average absolute deviation of 21.37K (4.67%) for a set of 2514 components, 14.46K (3.53%) for 2578 components for Stein and Brown, 13.22K (3.15%) for 2267 components for Constantinou and Gani, 10.23 (2.33%) for 1675 components for Marrero and Pardillo and 8.18K (1.90%) for 2766 components for Cordes and Rarey. This implies that the proposed method yielded the lowest average deviation with the broadest range of applicability. Also, on an analysis of the probability of prediction failure, only 3% of the data was greater than 20K for the proposed method. This detailed comparison serves as a very valuable tool for the estimation of prediction reliability and probable error. Structural groups were defined in a standardized form and the fragmentation of the molecular structures was performed by an automatic procedure to eliminate any arbitrary assumptions.

Preface

The work presented in this thesis was performed at the University of Natal, Durban from January 2002 to Januray 2004. The work was supervised by Professor D. Ramjugernath, Dr. J. Rarey and Professor J.D. Raal.

This thesis is presented as the full requirement for the degree of M.Sc. in Chemical Engineering. All the work presented in this thesis is original unless otherwise stated and has not (in whole or part) been submitted previously to any tertiary institute as part of a degree.

Y. Nannoolal (983186114)

As the candidate's supervisor, I, Prof. D. Ramjugernath, have approved this thesis for submission.

Prof. D. Ramjugernath

Acknowledgments

I would like to acknowledge the following people for their contribution to this work:

- Firstly, my supervisors, Professor D. Ramjugernath, Dr. J. Rarey and Professor J.D. Raal, for their support, ideas, wisdom and motivation during this work. Working with them has been a great learning experience in many fields.

- NRF International Science Liason, NRF-Thuthuka Programme, THRIP and University of Kwazulu-Natal Research Fund for financial support.

- On a personal note my parents, Suresh and Lalitha, my grandmother and late grandfather, my brother Dion and sister Anushka for years of support, guidance and motivation.

- My colleagues, Warren, Tyrone, Kruben, Hemraj, Pravesh, Thishen, Minal, Roger, Scott, Prathika, Ranjeetha, Alex, Bhaij and Natasha for their friendship and knowledge.

Table of Contents

Chapter Four

Chapter Five

Chapter Six

Chapter Seven

Chapter Eight

Appendix A

Appendix B

Appendix C

Appendix D

Appendix E

Appendix F

Appendix G

List of Figures

List of Tables

Nomenclature[1]

a, b, c, d, e, f	-	adjustable parameters
B	-	frequency of second-order group
C	-	first-order group contribution
D	-	second-order group contribution
E	-	third-order group contribution
G	-	Gibbs energy
H	-	enthalpy
M	-	molar mass
n	-	number of atoms
N	-	frequency of first-order group
O	-	frequency of third-order group
P	-	pressure
r	-	distance from centre of molecule
R	-	universal gas constant
S	-	entropy
T	-	absolute Temperature
V	-	volume

Greek letters

μ	-	chemical potential

Subscripts

b	-	boiling point

[1] All equations and variables are in SI units unless otherwise stated. Other symbols used are explained in context of this work.

est	-	estimated
trans	-	translation

Superscripts

sat	-	saturated
L	-	liquid
vap	-	vaporization
V	-	vapour

Chapter One

Introduction

The basis for any design and simulation of chemical, biochemical and environmental systems (for example, a chemical plant) is a reliable set of physical and chemical pure component and mixture properties. As it is not always possible to find experimental values in the literature and since measurement is expensive and time consuming or sometimes even difficult or impossible, estimation methods are generally of great value.

With the wide availability of computers and software for the simulation of chemical processes and environmental simulations (for example, compartment models for the estimation of the distribution of chemicals in the environment), there is a great need for physical properties, especially vapour pressures of a large number of rather exotic compounds (by-products, trace impurities, additives in design production, etc), which are not easily available from literature or experiment.

Another evolving application for accurate physical property estimation methods is computer aided molecular design (CAMD), which is focused on generating molecular structures for components with specific properties (vapour pressure, boiling point, viscosity, polarity, etc). During the optimization process, the computer will generate a large number of structures, for which experimental data are not available and the program has to rely on the accuracy of the predictive methods employed for verification.

Modern sophisticated process simulations employ physical property correlations for the estimation of organic compounds. However, a proper understanding of the thermodynamic assumptions underlying these simulators is needed to ensure proper application. Agarwal et al. (2001a, 2001b) recently published a paper entitled "Uncovering the Realities of Simulation". The paper proposes a number of examples suggesting that running sophisticated process simulations does not always guarantee correct results. Of these examples, Moura and Carneiro (1991) describe a problem where a commercial simulator was used for the evaluation of a 1,3-butadiene purification tower. It is well known that 1,2-butadiene is less volatile than 1,3-butadiene and would leave mostly through the tower bottoms. However, the simulator predicted that 1,2-butadiene would leave through the top. This is a quite a simple system for which experimental data and reasonable thermodynamic models are available.

The error was produced as a result of the simulator using critical properties predicted by the Cavett correlation (Cavett (1962)). This correlation incorporates the normal boiling point as the only input parameter and, as a result, the poor prediction was observed. If the physical properties used by the simulator were tabulated by the properties recommended by AIChE's Design Institute for Physical Property Data (DIPPR) and Dortmund Data Bank (DDB), it is evident that the acentric factor was incorrect (Table 1-1).

Table 1-1: Physical properties for 1,2-butadiene.

Physical Property	**Cavett (1962)**	**DDB (1973)**	**DIPPR (1969)**
Critical Temperature (°C)	184.72	170.55	170.67
Critical Pressure (KPa)	4065.8	4498.3	4500.2
Acentric Factor	0.0987	0.2550	0.2509

In principle, properties of a pure component can be derived from the structure of the molecule and, in some cases, state variables such as temperature and pressure. For the property estimation of pure components, group-contributions methods have been widely used. The properties of a compound are calculated as a function of structurally dependent parameters, which can be determined by summing the number frequency of each group multiplied by its contribution, on the assumption that the effects of the individual groups are additive. These methods have the advantage of supplying quick estimates with only the structure of the component known. Failure of various systems and simulations can often be traced to the questionable reliability and accuracy of the estimation methods for pure components.

There are a large number of group contribution methods for the estimation of physical properties, in particular the normal boiling point, available in literature. These methods will be described in the following chapter. In summary, current methods cannot provide a simple and accurate estimation of the normal boiling point across all chemical classes. Most methods have high average absolute deviations and fail drastically in the estimation of multifunctional compounds.

The major objective of this work is to develop a reliable group contribution estimation method for the prediction of the normal boiling points of non-electrolyte organic compounds. The term 'reliable' is significant in this context, since, the aim would be to develop a method where the probability of prediction failure is at its minimum. In order to do this, the analysis must be performed on a functional basis. The first step would be, however, to review current group contribution methods. These steps and the proposed development will all be discussed in the following chapters.

Chapter Two

Literature Review

2.1 Introduction

The definition of the boiling point is "the temperature at which the vapour pressure of a liquid is equal to the external pressure". The normal boiling point is the temperature at which the liquid boils when the external pressure is one atmosphere (101.325 KPa). Pure chemicals have a unique boiling point; mixtures on the other hand have a boiling point range. The boiling point is a function of temperature, i.e. the vapour pressure curve, of pure components is one of the most important properties for the calculation of many mixtures. It can be estimated from the vapour pressure value at one given temperature (for example, the normal boiling point) and the heat of vaporization.

The significance of a pure component boiling point, either estimated or measured, is that it defines the fugacity of a pure fluid at a given temperature. Also, the boiling point relates to the volatility of a chemical. For example, a distillation column, a key unit operation in separation technology is designed based on the relative volatility of the components. The boiling point also serves as input to models for the estimation of vapour pressure as a function of temperature. A more detailed description of the thermodynamics of the boiling point can be found in Chapter Three.

Group contribution is one of the simplest forms of estimation for any desired property, since it only requires the knowledge of the molecular structure. These methods are widely used for the synthesis and design of separation processes of industrial interest (for example, UNIFAC and modified UNIFAC for the prediction of real mixture behaviour).

The boiling point is also associated with molecular properties and molecular descriptors from molecular modelling. These properties, such as dipole moment, polarisability, hydrogen bonding, acid/base behaviour, etc, have a major effect on the boiling point. Therefore, these properties should be included in more sophisticated estimation methods. However, most of these properties can only be obtained from molecular simulation. This requires more complex calculations than simple group contribution techniques.

Thus, in numerous cases the most accurate method is not the most convenient to use. In general, an approximate estimate which is simple and can be hand calculated is generally preferred to complex methods requiring more than the molecular structure as input.

The estimation methods considered here are those that use only the structural information of the molecule. Since, if the boiling point is not available, it is likely that other properties (for example, critical temperature) are not available as well.

2.2 Overview of Available Group Contribution methods

There are several reviews on property estimation methods available; the most popular are probably those by Reid et al. (1987). There are more detailed reviews by Horvath (1992) and Boethling and Mackay (2000), and a more recent one by Poling et al. (2001), which provides a brief description of the methods. However, these reviews show that many of the estimation methods available have been derived for a specific homologous series; chemical classes such as hydrocarbons or alcohols, etc. Within such a class, boiling point estimation can be fairly accurate, however, because of their limited applicability, these methods will not be considered here.

No existing method can provide a simple and accurate estimation of boiling point across all chemical classes. Most general methods have average absolute deviations between 10-30 ^{o}C when dealing with compounds with just a single functional group. Consequently, in the estimation of multifunctional compounds, the methods often fail drastically, with extremely high deviations occurring.

The focus of this review covers group contribution estimation methods for the prediction of normal boiling points not limited to individual chemical classes. Table 2-1 presents an overview of these methods, which probably represent the best methods for the estimation of normal boiling points from group contribution thus far. A brief description of these methods will be given. These methods will later provide references, as a comparison to the proposed method.

Table 2-1: Overview of boiling point estimation methods provided in this chapter.

Author	No. of Groups	Description	AAE
Joback and Reid (1987)	41	Linear model based on a set of 438 components	12.9
Constantinou and Gani (1994a)	First order – 78 Second order - 42	Exponential model based on a set of 392 components. Second order contributions are based upon conjugation effects.	5.4
Marrero and Pardillo (1999)	165	Model including molecular weight based on a set of 507 components. Bond contributions are now used.	4.9
Stein and Brown (1994)	90	Linear model, however with two temperature correction models based on a set of 4426 components.	15.5
Cordes and Rarey (2002)	First order – 86 Second order - 7	Model including number of atoms based on a set of 2550 components. Second order group corrections also used.	-
Marrero and Gani (2001)	First order – 182 Second order – 123 Third order - 66	Exponential model based on a set of 1794 components. With second and third-order corrections.	5.89

(AAE - Average Absolute Deviation (K) given by the respective authors)

2.2.1 Joback and Reid (1987)

Joback and Reid examined many different types of estimation equations requiring group-contributions and selected Equation 2-1 for the prediction of the normal boiling point. This equation employs a linear relationship between the boiling point and the sum of group increments. They assumed no interaction between groups and structurally-dependant parameters are thereby determined by summing the number frequency of each group multiplied by its contribution. This linear relationship is only valid within a certain range of boiling points. This significantly limits the range of applicability (approximately T_b = 300 to 500K).

$$T_b = 198.2 + \Sigma N_i C_i \qquad (2\text{-}1)$$

They also employ only 41 molecular groups, which oversimplifies the molecular structure thus making isomers indistinguishable. Overall this is insufficient to capture the structural effects of organic molecules and is the main reason for the poor accuracy of the method. Table A-1 presents the 41 structural groups and their respective contributions. These groups are similar to Lydersen (1955) with the omission of >Si< and >B-, but with the inclusion of =N-(ring).

Multiple linear regression techniques were carried out on a set of 438 components to determine the group contributions for each structurally-dependant parameter. In the regression procedure, optimum values are generally obtained by minimizing the sum of squares of the absolute errors determined by the difference between the estimated and experimental property values. However, Joback and Reid suggested that minimizing the sum-of-squares of the errors weighted outliers too heavily, thus the sum of absolute errors was chosen. They employed a rather limited number of experimental boiling points compared to some other methods. This led to slightly higher errors for such outliers but provides an improved estimation procedure for the majority of compounds. To an engineer, in design and simulation of chemical systems, this would not be appropriate as the probability of the method failing is now higher. In general, the sum of squared errors is employed as this will lead to a better distribution of the predicted values.

The advantage of the method is in its simplicity; however, the relatively small range of compounds and poor predictions leads to the downfall of the method. Joback reported an average absolute deviation of 12.9K (3.6%) for the above data set. However, on a set of 2506 components obtained from the Dortmund Data Bank (DDB), an average deviation of 21.4K was obtained. Many authors have, however, followed up the work of Joback and Reid making use of it as a starting point.

2.2.2 Constantinou and Gani (1994a)

Second-order or second level approximations have the effect of differentiating amongst isomers. The basic premise is to provide enough information about the molecular structure of the compound, such that a significantly improved prediction of properties can be made. Constantinou et al. (1993, 1994b) provided an additive property estimation method, which is based on conjugate operators and applicable to organic compounds. However, the generation of conjugate forms is a non-trivial issue and requires a symbolic computing environment. Constantinou and Gani (1994a) applied the method of Constantinou et al. (1993, 1994b) based on second order conjugate forms to group contributions. The method proposed a property estimation, which is performed at two levels. The basic level has contributions from first-order functional groups and the next level has second-order groups, which have the first-order groups as building blocks. Thus, their method allows for both a first-order approximation (using first-order groups) and a more accurate second-order approximation (using both first- and second-order groups).

They had considered group contribution-based computational tools, which needs to accommodate two separate first-order molecular-structure descriptions, one for the prediction of pure component properties (Reid et al. (1987), Lyman et al. (1990)) and another for mixture property estimations (Fredenslund et al. (1977), Derr and Deal, (1969)). To circumvent this drawback, they proposed to use as first-order groups, Table A-2, the set of groups commonly used for the estimation of mixture properties. The disadvantage of this selection is that a group appearing in an aliphatic ring is considered equivalent to its identical non-ring one. Also, another important disadvantage of the group definition is that there is no theoretical identification. Therefore each group has a single contribution independent of the type of compound involved. There were 78 first order groups, quite similar to those used by Joback and Reid; most of the new groups being sub-divisions and quite a few of them being redundant as well.

Since their estimation was primarily based upon information about the molecular structure only, the idea was to include a different level of approximation. Thus Constantinou and Gani introduced second-order groups to provide more structural information about the compound. Their ultimate goal was to enhance the accuracy, reliability and the range of applicability of the property estimation, and overcome proximity effects and isomer differences. Contrary to first-order groups, there can be molecular structures, which do not need any second-order groups. The definition and identification of second-order groups, however, must have a theoretical basis. Thus, they proposed the principle of conjugation, as introduced by Constantinou et al. (1993, 1994c).

The theoretical background to conjugation is that compounds are represented as hybrids of many conjugates. Each conjugate form is an idealized structure with integer-order-localized bonds and integer charges on atoms. The purely covalent conjugate form is the dominant conjugate and the ionic forms are the recessive conjugates, which can be obtained from the dominant form by re-arrangement of electron pairs. A conjugation operator defines a particular pattern of electron arrangement. When applied to the dominant conjugate, an operator yields an entire class of recessive conjugates. Conjugation operators are represented by a distinct sub-chain with two or three bonds, such as C-C-C-H and O=C-C. Figure 2-1 presents a dominant conjugate, a generated recessive conjugate and the corresponding conjugation operator.

In the framework, the properties are estimated by determining and combining properties from its conjugate forms. Properties of conjugate forms are estimated through conjugation operators. In the method, they took the following as the principles for the identification of second-order groups:

- The structure of a second-order group should incorporate the distinct sub-chain of at least one important conjugation operator.
- The structure of a functional second-order group should have adjacent first-order groups as building blocks and it should be as small as possible.
- Second-order groups based on common operators(s) should be equally treated in the method.
- The performance of second-order groups is independent of the molecule in which the group occurs, satisfying a fundamental group-contribution principle.

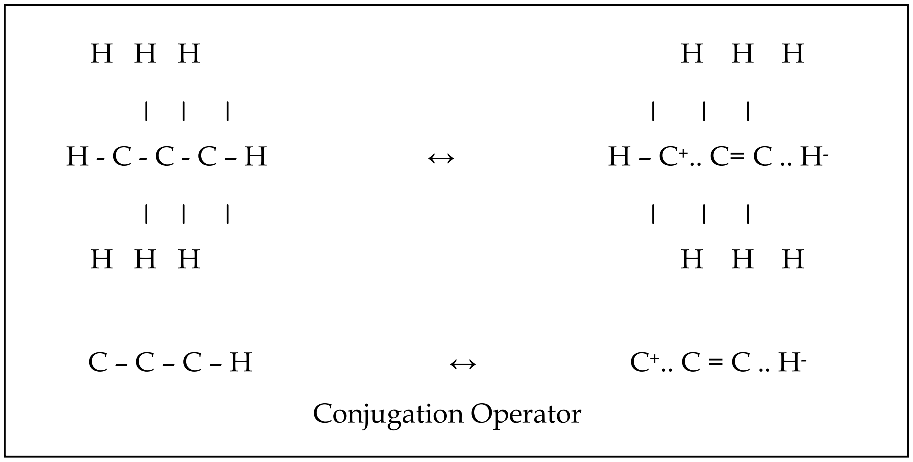


Figure 2-1: Dominant, recessive conjugates and conjugation operator

Table A-3 lists second-order groups that have been defined for the method and their contributions. The idea of conjugation is primarily based on the recessive conjugate proposing another form of the molecule. Thus in the property estimation, the molecule is now a mixture of dominant and recessive conjugates. The second-order groups account for the alternate form, or recessive conjugates. However, in many cases the possibility of a recessive conjugate form existing at atmospheric conditions is almost zero. For example, in Figure 2-1, the molecular structure of propane is presented. Propane is a non-polar covalent hydrocarbon with sp^3 carbon atoms, and the possibility of a recessive conjugate existing at atmospheric conditions is essentially zero. This would mean that a second order-group would now be defined for propane, even though there isn't one. This second order group would be derived for other components, of which for most of these components, the recessive conjugate form does not exist. Thus, under certain circumstances, the same molecule may be described in different ways because of the over-complication of this method. Now the aim would be to view the group's importance from its scientific and mathematical significance. Thus, mathematically, the group will consider components where the form does not exist and fit a contribution to coincide with these components. Scientifically, the significance of the group has a relatively small influence on the boiling point. The method suggested a logarithmic model equation for the boiling point estimation, Equation 2-2.

$$T_b = 204.359 \ln (\sum_i N_i C_i + W \sum_j B_j D_j) \tag{2-2}$$

The constant W is assigned a value of zero for a first-order approximation and unity in the second-order approximation, where both first and second-order group contributions are involved. 392 experimental data points were used in the regression. After the selection of data, a least square analysis had been carried out to determine the contributions of first- and second-order groups (adjustable parameters). Constantinou and Gani reported an average absolute error of 5.35K (1.42%) for the above data set, however, on a data set of 2259 components from the DDB, the average absolute error was 13.3K.

2.2.3 Marrero and Pardillo (1999)

Estimations of the normal boiling point have a strong dependence on the actual conformation of the molecule. This also affects the critical constants of the compound, indirectly, due to their dependence on the normal boiling point. To overcome the above limitation, Pardillo and Gonzalez-Rubio (1997) had first proposed a new structural approach called Group Interaction Contribution (GIC), which considers the contribution of interactions between bonding groups instead of the contribution of simple groups. Based on the above approach (GIC), Marrero and Pardillo (1999) proposed a new method, which estimates the boiling points and critical constants of pure organic compounds.

Marrero and Pardillo selected 39 simple groups, which can also be referred to as first-order groups, to generate a consistent set of group-interactions that allows one to treat a wide variety of organic compounds. These groups are similar to Joback and Reid, presented earlier, with the omission of =NH and =N-(non-ring). The model equation is also similar to the one proposed by Joback and Reid, Equation 2-4. In addition, they proposed a new alternative non-linear equation for estimating the boiling point, which involves the molar mass of the molecule, Equation 2-3.

$$T_b^* = M^{-0.404} \sum N_i\, C_i + 156 \tag{2-3}$$

$$T_b = 204.66 + \Sigma N_i\, C_i \tag{2-4}$$

The contributions of the group-interactions for Equation 2-3 (T_b^*), and Equation 2-4 (T_b) are presented in Table A-4. The group-interaction proposed here should actually be known as, and from now on referred to as, bond contributions. Because there is no physical interaction between groups, rather it's just the bonding between two defined groups. They did not calculate some bond-contributions because of the lack of property values for the compounds involved in these interactions. Also groups that were used to derive the bond contributions were from the Joback and Reid method, where the range of applicability is small and groups were poorly defined.

They employed the singular-value decomposition procedure (Forsythe et al., 1977) as the optimization algorithm for linear regression. For non-linear regression, they used the well-known Levenberg-Marquadt procedure.

On a data set of about 2800 components from the DDB, only 1665 components were fragmented for the above-mentioned method. Thus, despite the advantages of the method, their ranges of applicability are still quite restricted. Due to the relatively small data sets used in the development of these methods, which usually includes about a few hundred relatively simple compounds, their predictive capability usually breaks down when dealing with large, polycyclic or poly-functional compounds. The bond contributions do provide a better estimation for isomers; however, as with Constantinou and Gani, their physical significance to physical properties is minimal. For Equation 2-3 and 2-4, an average absolute deviation of 4.87K and 6.36K was reported, respectively, on a data set of 407 components. However, an average absolute deviation of 10.3K was obtained on the data set from the DDB, for Equation 2-3.

2.2.4 Stein and Brown (1994)

Stein and Brown (1994) proposed a new estimation method for the boiling point which is an extension of the Joback method. This extension is mainly the increase in number of groups from 41 to 85. However, many of the new groups are just subdivisions of those Joback and Reid used, where now the molecular groups contains C, N, O, S, halogens, 3 P groups, 3 Si groups and one each for B, Se and Sn. These groups were derived by evaluating 4426 compounds. Table A-5 presents the groups with their regressed values.

Following from the method of Joback and Reid, Stein and Brown also used a similar linear model for the estimation, Equation 2-5. However, on the larger data set they found the higher boiling compounds did not fit the linear model, which tended to over-predict the normal boiling point. Thus they proposed a boiling point model temperature correction based on the error deviation obtained from Equation 2-5, which is Equation 2-6 and 2-7 for a prediction of less than or equal to 700K and greater than 700K respectively.

$$T_b = 198.2 + \sum_i N_i C_i \qquad (2\text{-}5)$$

$$T_b\,(\mathrm{corr}) = T_b - 94.84 + 0.5577\,T_b - 0.0007705\,T_b \qquad \text{For } T_b \leq 700\text{ K} \qquad (2\text{-}6)$$

$$T_b\,(\mathrm{corr}) = T_b + 282.7 - 0.5209\,T_b \qquad \text{For } T_b > 700\text{ K} \qquad (2\text{-}7)$$

Together with the Joback and Reid method, the Stein and Brown method assumes no interaction between groups. However, the group definition changes if the fragment is in a ring or in a defined structural position, for example, on a secondary carbon. This emphasises the classification of structural groups for a more accurate prediction. For the above data set, the method had an average absolute deviation of 15.5K (3.2%). Stein and Brown also tested their method on an independent test set of 6584 components and found an average absolute deviation of 20.4K (4.3%). For 2579 components obtained from the DDB, an average absolute deviation of 14.5K was obtained.

2.2.5 Marrero and Gani (2001)

Marrero and Gani (2001) proposed a new group-contribution method that allows an accurate and reliable estimation for a wide range of compounds, including large and complex compounds. In their method, there are now three levels of approximation.

The first level has a large set of simple groups that is able to partially capture proximity effects, but is unable to distinguish between isomers. The groups are also similar to other first order groups stated previously. For this reason, the first level of estimation is intended to deal with simple and mono-functional compounds. Marrero and Gani assumed the following criteria for the description of first order groups:

- The set of groups should allow the representation of a wide variety of chemical classes.
- Each group should be as small as possible because very large groups are generally not desirable.
- A detailed first-order approximation of aromatic compounds should be provided at a first level of estimation; groups in the form a[2]C-R, such as aC-CO. Also, two specific groups have been included, aN and aCH, for the representation of pyridines and nitrogen-containing aromatics. Furthermore, three different corrections have been included of the form aC, to differentiate among, (a) carbon atoms shared by different aromatic rings in a fused system, (b) carbon atoms shared by both aromatic and non-aromatic rings in a fused system and (c) any other substituted aromatic carbon that does not fall into the above category.
- The set of first-order groups should allow the distinction between groups occurring in cyclic and acyclic structures. It was found that better property estimation is achieved by using separate ring and non-ring groups for cyclic and acyclic structures.
- First-order groups should describe the entire molecule. In other words, there should be no fragment of a given molecule that cannot be represented by first-order groups. It should also be noted that no atom of a given molecule can be included in more than one group.
- The contribution of any first-order group should be independent of the molecule in which the group occurs, which satisfies one of the fundamental principles of the group-contribution approach.

Based upon the above criteria of identification, a comprehensive set of first-order groups has been identified and are presented in Table A-6. It should be noted that some rules have to be followed in order to correctly assign the groups that occur in a given compound. It is assumed that heavier groups hold more information about the molecular structure than lighter groups; consequently, the golden rule is that, if the

[2] Aromatic

same fragment of a given compound is related to more than one group, the heavier group must be chosen to represent it. There are only two exceptions to this rule; one is in the case of aromatic substituents for which groups of the form aC-R must be used. The other exception occurs for ureas and amides, for which special functional groups are provided.

The second level permits a better description of poly-functional compounds and differentiation amongst isomers. The following criteria have been considered for the identification of second-order groups:

- There can be compounds that do not need any second-order contribution.
- Also, the entire molecule does not need to be described by second-order groups. Second-order groups intend to describe molecular fragments that could not be adequately described by first-order groups, and thereby yielded a poor estimation at the first level.
- As is has been suggested, the main purpose of second-order groups is to differentiate among isomers. Accordingly, specific groups are provided with this objective in mind. These groups allow differentiation not only in alkanes, alkenes and other open-chain structures, but also in aromatic compounds for which special groups such as AROMRING[3]s^1s^2, etc., have been included.
- Second-order groups should be allowed to overlap each other. That is, a specific atom of the molecule may be included in more than one group. It is necessary to prevent a situation in which one group overlaps completely with another group, since it would lead to a redundant description of the same molecular fragment. The contribution of any group should be equal in whichever molecule the group occurs.

Second-order groups are, however, unable to provide a good representation of compounds containing more than one ring as well as, in some cases, open-chain poly-functional compounds with more than four carbon atoms in the main chain. Thus, for this reason, a further level is required to provide a better description for these types of compounds. Second-order groups are presented in Table A-7.

[3] Aromatic ring

Thus, the method proposed third-order groups, which intend to represent the molecule at the third level of approximation. The third level of estimation allows estimation of complex heterocyclic and large (C= 7-60) poly-functional acyclic compounds. The criteria used for the identification of third-order groups are analogous to those used for second-order groups. Third-order groups are presented in Table A-8.

The property-estimation model has the form of the following equation:

$$T_b = 222.543 \ln\left(\sum_i N_i C_i + w \sum_j B_j D_j + z \sum_k O_k E_k \right) \tag{2-8}$$

The determination of the adjustable parameters for the models has been divided into a three-step regression procedure.

- Determination of contribution C_i of the first-order groups while w and z are set to zero.
- Then, w is set to unity and z to zero and another regression is carried out using the previous C_i to determine the contribution $D_{j.}$
- Finally, both w and z set to unity, and the contribution E_k determined using previous contributions.

This stepped regression scheme ensures independence among contributions of the three levels of approximation. The optimization algorithm used for data fitting was the Levenberg-Marquadt technique. The experimental data used in the regression has been obtained from the Computer Aided Process Engineering Centre (CAPEC-DTU) database.

Overall, the method is highly complex, incorporating an extremely large definition of first-order groups, 182, for a data set of only 1794 components. These groups are mainly subdivisions of their previous methods of which many of the groups are redundant. However, the method differentiated between groups in much more detail according to chain, rings structures etc. This plays a major role in the boiling point prediction, as will be seen in the Cordes and Rarey (2002) and the proposed method. They were also 122 second-order groups and 66 third-order groups. These groups should have a theoretical basis for their definition. However, this is not clear in the description of the method. These groups seem to be derived for components where there are extreme deviations. This fact can also be seen by the regression procedure described above. Thus again, the scientific significance of these groups are minimal, since there seems to be just the building up of structural groups due to the over-complication of the method. Thus, their predictive capability is questionable. The method reported an average absolute deviation of 7.90K (1.8%) for a first-order approximation, 6.38 (1.4%) for a second-order approximation and 5.89K (1.4%) for a third-order approximation. It can be seen that there is not much difference in average absolute deviations between the different levels of approximations mentioned above.

At this point, it should be noted that the proposed method will be compared to all the methods described in this section, excluding the above Marrero and Gani (2001). For the data set that will be used in the proposed method, all the results will be computer generated. However, the Marrero and Gani method is fairly new and also definitely the most complex, and has not been incorporated into the software as yet. There will be, however, a manual fragmentation done on n-alkanes as a comparison to the proposed method.

2.2.6 Cordes and Rarey (2002)

The method of Cordes and Rarey (2002) suggested a new approach for the prediction of the boiling point. Instead of improving the estimation by involving different levels of approximation, they proposed a more scientific definition of first-order groups, which forms the basis of group contribution. The groups were defined according to their chemical neighbourhood. Thus it became apparent that

- There is no need to distinguish between carbon and silicon as a neighbouring atom since, both elements has almost similar structural characteristics.
- Very electronegative (N, O, F and Cl) or aromatic neighbours often significantly influence the contribution of a structural group.
- It is usually of great importance whether a group is part of a chain, ring or aromatic system.

Tables A-9 and A-10 presents the 86 groups and 7 corrections proposed by the method respectively, and their contributions.

Another complexity of group contribution is the fragmentation of molecules into their respective groups. By hand this can be a time consuming and tedious procedure. However, Cordes and Rarey proposed an automatic fragmentation algorithm, for which structural groups are defined in a standard form, and the automatic procedure performs the fragmentation. Thus for all methods this can be easily done, and a comparison on common sets of data can be easily obtained. This procedure will be explained in more detail later on.

It should be noted that in most of the available group contribution methods, important features such as the boiling point model and experimental database were not investigated in detail. Thus Cordes and Rarey also investigated these features and aimed at developing an improved expression for the dependence of the normal boiling point on the sum of group increments and a significantly larger set of reliable experimental information. This model is presented below.

$$T_b = \frac{\sum N_i C_i}{n^{0.6713} + 1.4442} + 59.344 \tag{2-9}$$

The expression provides a better description of the dependence of T_b on molecular size, as it carries the additional advantage, that via the number of atoms in the molecule, an additional and readily available quantity more or less independent from the sum of the increments is introduced. Normal boiling point data for approximately 2800 components are available on the Dortmund Data Bank (DDB), out of which the structural groups were constructed from 2550 components.

They also decided to compare their method as per functional groups, for example, alkanes, ethers etc, and in most cases were far better than the above mentioned methods. Since the proposed method of this thesis will follow the work of Cordes and Rarey, a detailed comparison of the average absolute deviations of each functional group will not be done here. An important disadvantage of the method, however, is the inability to differentiate amongst isomers, and this will be addressed in the proposed method.

From all the above methods, properties of large, complex and multi-functional compounds of interest in biochemical and environmental studies cannot be accurately estimated using the available methods. Neither of the methods has investigated the thermodynamics, general and physical chemistry surrounding the boiling point, or any other physical property of that matter. Rather, previous methods have provided more information about the molecule, which has little or no relevance to the boiling point. Importantly, the physical interaction between molecules has a large influence on the normal boiling point and will be investigated in Chapter 3.

2.3 Group Vector Space

An evolving application of group contribution is group vector space (GVS). This takes into consideration the specific position of the group in a molecule. Consider the methods of Wen and Qiang (2002a, 2002b) which employs this approach. Wen and Qiang (2002a) suggested a GVS approach for hydrocarbons and Wen and Qiang (2002b) for organic compounds. Because of the range of applicability, the latter method will be discussed here.

The method selected 40 simple groups to describe organic compounds. These groups are the same as those used by Joback and Reid. The molecule is considered to be in a given space, and every group in the molecule is only a point in the space. For convenience, since there are graphs with different number of points, these graphs are all expressed as graphs with five points. Consequently, an organic molecule can be expressed as seven topologic graphs (Figure 2-2).

Considering the chain graph first, the dimension number of the space is equal to the number of end points on the chain, and one end point has determined a dimension of the space. The coordinate of an end point in the dimension determined by it is zero, while the coordinate of another point in this dimension is the distance from that point to the end point. For the cyclic graph, one ring represents a dimension. In that dimension the coordinate of the ring point equals the number of points on the ring, and the coordinate of the non-ring point equals the sum of the distance from the point to ring and the number of points on the ring. If the route from the ring point to the end point is not unique, the shortest route should be selected. So, the dimension number m of the space for a graph is equal to the sum of the number k_e of end points and the number k_r of rings in the graph. Every point in the graph has *m* coordinates in the *m*-dimensional space. The graph may be described by a space matrix, where the number of rows in the matrix equals the number of points in the graph and the number of columns equals the dimension number of the space. The space matrices of the above seven topologic graphs are presented in Figure 2-3.

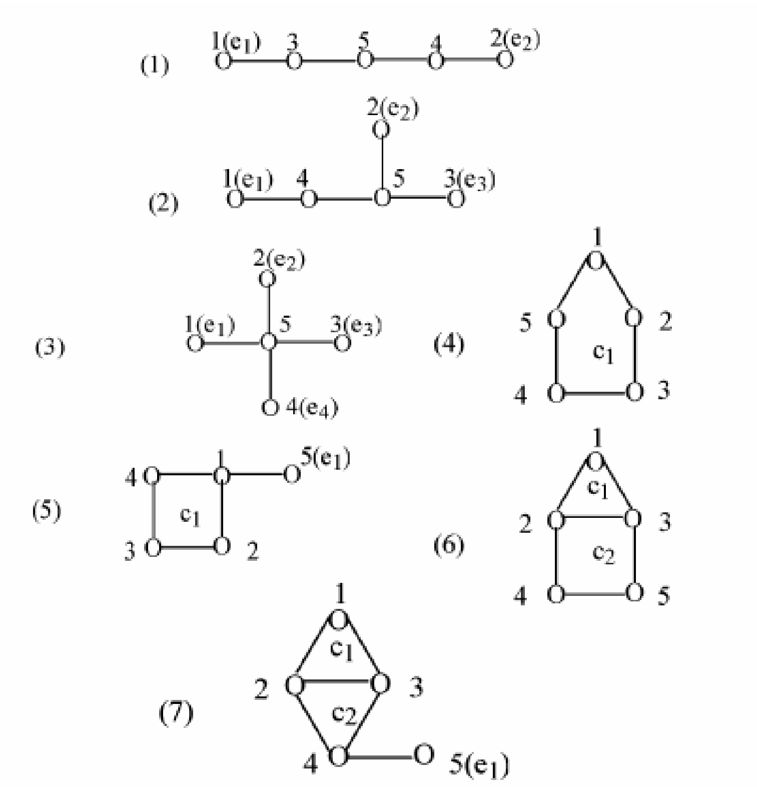


Figure 2-2: Organic molecules expressed as seven topological graphs.

The matrices show that the space position of the point *i* in the graph can be represented by an m-dimensional vector (b_{i1}, b_{i2}, …, b_{im}). Thus, the module α_i can be determined by Equation 2-10.

$$\alpha_i = \left(\sum_{j=1}^{m} b_{ij}^2 \right)^{1/2} \quad (i = 1 - 5) \qquad (2\text{-}10)$$

The average square root of the module of some point *i* is defined as the module index v_i of this point vector (Equation 2-11). The quantity v_i is used to describe the point *i* position in the space. In this analogy, the module index v_i of group *i* in the molecule is taken to characterize the position of that group in the molecular space. Thus, every simple group, except halogen groups, has its own independent module index. For the four halogen groups, their module indexes were determined to be the same as those of the hydrocarbon groups with which they were connected.

$$
(1)\ \begin{array}{cc} & \begin{matrix} e_1 & e_2 \end{matrix} \\ \begin{matrix}1\\2\\3\\4\\5\end{matrix} & \begin{bmatrix} 0 & 4 \\ 4 & 0 \\ 1 & 3 \\ 3 & 1 \\ 2 & 2 \end{bmatrix} \end{array};\quad
(2)\ \begin{array}{cc} & \begin{matrix} e_1 & e_2 & e_3 \end{matrix} \\ \begin{matrix}1\\2\\3\\4\\5\end{matrix} & \begin{bmatrix} 0 & 3 & 3 \\ 3 & 0 & 2 \\ 3 & 2 & 0 \\ 1 & 2 & 2 \\ 2 & 1 & 1 \end{bmatrix} \end{array};\quad
(3)\ \begin{array}{cc} & \begin{matrix} e_1 & e_2 & e_3 & e_4 \end{matrix} \\ \begin{matrix}1\\2\\3\\4\\5\end{matrix} & \begin{bmatrix} 0 & 2 & 2 & 2 \\ 2 & 0 & 2 & 2 \\ 2 & 2 & 0 & 2 \\ 2 & 2 & 2 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \end{array};
$$

$$
(4)\ \begin{array}{cc} & \begin{matrix} c_1 \end{matrix} \\ \begin{matrix}1\\2\\3\\4\\5\end{matrix} & \begin{bmatrix} 5 \\ 5 \\ 5 \\ 5 \\ 5 \end{bmatrix} \end{array};\quad
(5)\ \begin{array}{cc} & \begin{matrix} c_1 & e_1 \end{matrix} \\ \begin{matrix}1\\2\\3\\4\\5\end{matrix} & \begin{bmatrix} 4 & 1 \\ 4 & 2 \\ 4 & 3 \\ 4 & 2 \\ 5 & 0 \end{bmatrix} \end{array};\quad
(6)\ \begin{array}{cc} & \begin{matrix} c_1 & c_2 \end{matrix} \\ \begin{matrix}1\\2\\3\\4\\5\end{matrix} & \begin{bmatrix} 3 & 5 \\ 3 & 4 \\ 3 & 4 \\ 4 & 4 \\ 4 & 4 \end{bmatrix} \end{array};
$$

$$
(7)\ \begin{array}{cc} & \begin{matrix} c_1 & c_2 & e_1 \end{matrix} \\ \begin{matrix}1\\2\\3\\4\\5\end{matrix} & \begin{bmatrix} 3 & 4 & 3 \\ 3 & 3 & 2 \\ 3 & 3 & 2 \\ 4 & 3 & 1 \\ 5 & 4 & 0 \end{bmatrix} \end{array}
$$

Figure 2-3: Space matrices of the above seven topological graphs.

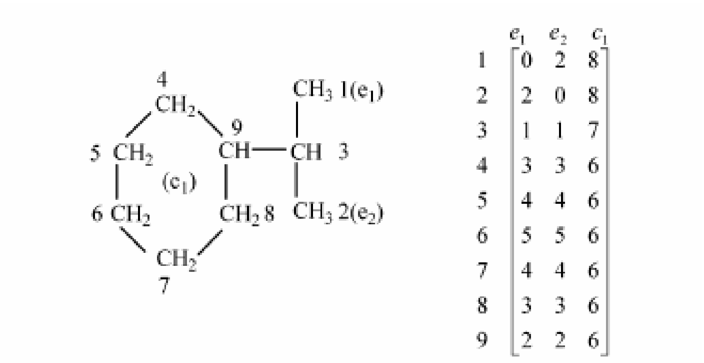
$$
\mathrm{v}_{\mathrm{i}} = \alpha_{\mathrm{i}} \Bigg/ \left(\sum_{\mathrm{j}=1}^{5} \alpha_{\mathrm{j}}^{2} \right)^{1/2} \quad (\mathrm{i} = 1 - 5) \tag{2-11}
$$

The normal boiling point, T_b, can be expressed by Equation 2-12. This expression incorporates a position contribution, ΔT_{bPi}, an independent contribution, ΔT_{bIi}, and a constant, ΔT_{b0i}. To improve the estimation accuracy, Wen and Qiang implemented a trial computation to obtain the optimum power index of T_b. The model and contributions were based on a set of 669 components.

$$
\mathrm{T}_{\mathrm{b}}^{1.5} = 1304.13 + \sum_{\mathrm{i}} \left(\sum_{\mathrm{j}=1}^{\mathrm{n_i}} \mathrm{v}_{\mathrm{j}} \Delta \mathrm{T}_{\mathrm{bPi}} + \mathrm{n}_{\mathrm{i}} \Delta \mathrm{T}_{\mathrm{bIi}} + \Delta \mathrm{T}_{\mathrm{b0i}} \right) \tag{2-12}
$$

The contributions for the above model are not presented in this work, since only the description of GVS is needed. GVS is a non-trivial issue and is a computational burden due to its complexity. For this reason and also since the method is fairly new, the method has not yet been incorporated into the DDB Artist program, and a comparison to the proposed method cannot be made. However, a comparison to previous methods, as reported by Wen and Qiang, is presented in Table 2-2. By the introduction of GVS, a far more accurate estimation is achieved as compared to the parent method proposed by Joback and Reid. The results are, however, slightly less accurate than Constantinou and Gani, but with a larger set of data. Thus, a comparison to the proposed method can be achieved by assuming an average deviation similar to that of Constantinou and Gani.

The derivation of GVS is relatively complex, thus, an example for the estimation of the normal boiling point of isopropylcyclohexane is presented below:



$$\begin{array}{c|ccc} & e_1 & e_2 & c_1 \\ \hline 1 & 0 & 2 & 8 \\ 2 & 2 & 0 & 8 \\ 3 & 1 & 1 & 7 \\ 4 & 3 & 3 & 6 \\ 5 & 4 & 4 & 6 \\ 6 & 5 & 5 & 6 \\ 7 & 4 & 4 & 6 \\ 8 & 3 & 3 & 6 \\ 9 & 2 & 2 & 6 \end{array}$$

Figure 2-4: Module α_i and corresponding module index v_i for isopropylcyclohexane.

From Figure 2-4, the module α_i and corresponding module index v_i is calculated from Equations 2-10 and 2-11, respectively. The results are presented in Table 2-2.

Table 2-2: Values of α_i and v_i for isopropylcyclohexane.

Group no.	1	2	3	4	5	6	7	8	9
α_i^2	68	68	51	54	68	86	68	54	44
v_i	0.3482	0.3482	0.3015	0.3103	0.3482	0.3915	0.3482	0.3103	0.2801

From Table 2-2, the computation of $\Sigma\alpha_i^2$ is 561. Thus, the computation of group contributions and corresponding $\sum_{j=1}^{n_i} v_j$ and n_i are presented in Table 2-3. The group contributions (ΔT_{bP}, ΔT_{bI}, ΔT_{b0}) are not presented in this thesis, but are referenced to Wen and Qiang (2002a).

Table 2-3: Values of group contributions and corresponding $\sum_{j=1}^{n_i} v_j$ and n_i for isopropylcyclohexane.

Group i	v_i	n_i	ΔT_{bP} / 100	ΔT_{bI} / 100	ΔT_{b0} / 100
$-CH_3$	0.6964	2	7.969	2.209	0.224
>CH-	0.3015	1	1.483	11.312	-0.037
$(-CH_2-)_R$	1.7085	5	6.810	7.384	-6.829
$(-CH-)_R$	0.2081	1	2.768	10.577	-0.027

The normal boiling point can now be estimated from Equation 2-12 using the group contributions from Table 2-3. Thus, for the Wen and Qiang method, the estimated value for isopropylcyclohexane is 426.3 K. The experimental value is 427.7 K with a relative deviation of -0.33 %.

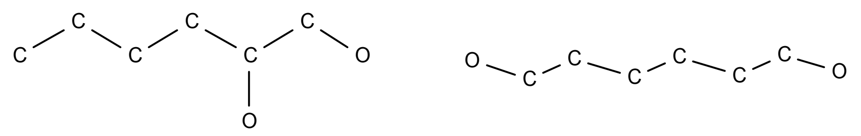
GVS accounts for the position of a specific group in a molecule. The general question is whether it is non-different from the method proposed by Cordes and Rarey. The latter method is, however, less complicated to derive. In theory, group contribution can capture the positioning of a single functional group by a more scientific group definition. However, it will be seen in the following chapters, that a specific functional group's position relative to the position of another functional group has a greater significance to the boiling point estimation, and this cannot be captured by group contribution. This can also be observed from the difference in normal boiling points of 1,2-hexanediol and 1,2-hexanediol (Figure 2-4), which is 53.6K. The major influence on the normal boiling point in this example is the dipole moment, which is a result of the different positioning of the alcohol groups relative to each other. Thus, GVS is not able to account for the relative positions of functional groups in a molecule and is also not able to identify the different groups in the estimation of the normal boiling point. In other words, GVS does not take into account the electronic properties of the functional groups, for example, electronegativity, in its calculation (see example above). However, the dipole moment is able to do this and is arguably less complicated to calculate from a molecular simulation package.

Table 2-4: Comparison of the estimation accuracy of different models as reported by Wen and Qiang (2002b).

Method	**NC^a**	**AAE^b**	**APE^c**
Joback and Reid	438	12.9	3.6
Marrero and Pardillo	507	6.48	1.73
Constantinou and Gani	392	5.35	1.42
Wen and Qiang	669	5.51	1.40

^{a} Number of components, ^{b} AAE – Average absolute error, ^{c} APE – Average percent error

Majority of group contribution methods available in literature employ the Joback and Reid method as the parent method. The poor definitions and results from the Joback and Reid method provide a poor base for further development. Furthermore, other methods which includes molecular descriptors as described in the review by Poling et al. (2001), and methods including quantitative structure-property relationship (QSPR) such as Ericksen et al. (2002), employ poor bases for their development. Consequently, one of the aims for the proposed method is to provide a convenient base, such that, further development which includes molecular descriptors, molecular properties, etc, will provide a significantly more accurate estimation.



1,2-hexanediol (T_b = 469.6K) 1,6-hexanediol (T_b = 523.2K)

Figure 2-5: Normal boiling points of 1,2-hexanediol and 1,5-hexanediol.

Chapter Three

Theoretical Considerations

3.1 Introduction

What influences the normal boiling point? This is a question that generally needs to be answered in order to develop an estimation method for the normal boiling point. Constantinou and Gani and Marrero and Pardillo suggested a conjugation and bond contribution approach, respectively. Cordes and Rarey defined the structural first-order groups according to its neighbouring atoms. Even though these methods and results were discussed in the preceding chapter, the general question is which of these approaches is more significant to the normal boiling point. Consequently, this chapter will try and answer these questions, indirectly, and analyze the different factors influencing the normal boiling point.

3.2 General Theory

The bubble point of a liquid is described as the temperature at which the first vapour forms. A liquid at its bubble point requires just enough energy, equal to the latent heat of vaporization, for a phase change from a liquid to vapour. From the second law of thermodynamics it follows that this phase change accompanies a positive entropy change. This proposes that the normal boiling point is interrelated to the enthalpy and entropy change at atmospheric conditions. This relationship needs to be sought out.

Consider the equilibrium between two phases in thermal and mechanical equilibrium (dT=0, dP=0) is achieved, when dn$_i$ (the transfer of n moles of component i) is equal to 0, for all components. An alternate form of the fundamental property relation involving the Gibbs function can be written as follows:

$$d(nG) = -(nS)dT + (nV)dP + \sum \mu_i \, dn_i \tag{3-1}$$

At constant temperature, pressure and mass of a system, this means that (from Equation 3-1):

$$d(nG) = 0 \qquad \text{and} \qquad G \equiv \text{minimum} \tag{3-2}$$

Thus, it can be easily shown that this leads to equal chemical potentials in all phases for each component (chemical equilibrium between the phases):

$$\Delta G^{\alpha\beta} = G^{\alpha} - G^{\beta} = 0 \tag{3-3}$$

Using the well known equation G=H-TS, the following relationship can be derived:

$$T = \frac{\Delta H^{\alpha\beta}}{\Delta S^{\alpha\beta}} \tag{3-4}$$

At atmospheric conditions, we have the relationship for the normal boiling point (T_b) as a ratio of ΔH^{sat} and ΔS^{sat} (this generalization can also be referred to as Trouton's rule):

$$T_b = \left(\frac{\Delta H^{sat}}{\Delta S^{Sat}} \right)_{1atm} \tag{3-5}$$

The enthalpy of vaporization is the difference between the enthalpy of the saturated vapour and that of the saturated liquid at the same temperature. Molecules in the vapour phase do not have the energy of attraction that those in the liquid have, therefore energy must be supplied for vaporization to occur. Thus, with increasing attractive forces, as can be seen by more polar molecules, this accompanies a higher enthalpy of vaporization. At the normal boiling point, the total interaction between the molecules in the vapour phase is small as compared to that in the liquid phase. Consequently, the enthalpy of vaporization can be approximated by the total intermolecular interaction in the liquid phase.

Entropy corresponds to the number of arrangements (positions and/or energy levels) that are available to a system in a given state. The more ways a particular state can be achieved, the greater is the likelihood (probability) that state will occur. (Nature spontaneously proceeds towards the states that have the highest probabilities of existing).

Consider the following theory involving the molecular basis of entropy, cited by Barrow (1985). "The equilibrium of A and B in which B has the higher entropy, for example, can be understood in terms of the fact that for some reason there are more available quantum states corresponding to B. There are therefore more ways of distributing the atoms in these states so that a molecule of type B is formed than there are ways of arranging the atoms in the quantum states so that a molecule of type A is formed. The tendency of A to change over to B, even if no energy driving force exists, is therefore understood to be due to the driving force that takes the system from a state of lower probability, i.e., of few quantum states and a few possible arrangements, to one of higher probability, i.e., one of many available quantum states and more possible arrangements. The qualitative result from this discussion is: *A substance for which the molecules have more available quantum states has the higher probability and therefore the higher entropy."*

"The molecular explanation of the entropy change in a process is basically quite simple. In practice, of course, it is now always easy to see whether a process, or reaction, produces a system with more, or less, available quantum states or energy levels. Thus, for the liquid-to-vapour transition a large entropy change increase occurs. The difficulties encountered in a molecular understanding of the liquid state make it very difficult to evaluate this entropy increase from the molecular model."

The description of the molecular motions can be ascertained from its contribution of translation, vibration and rotation. Thus, it is important to understand the molecular basis of entropy. From understanding these molecular motions, a quantitative value for thermodynamic functions can be calculated. It can also be shown that values calculated for the entropy of an ideal gas agrees with thermodynamic third law values.

Amazingly most liquids have about the same molar entropy of vaporization. Following

Trouton's rule, the molar entropy increases by the same amount when 1 mol of any substance is changed from liquid to vapour at its normal boiling point. With only a few exceptions, the entropy of vaporization is approximately constant, 88 ± 5 $J.mole^{-1} K^{-1}$. There are two exceptions to the above rule. The first exception is for components with low boiling points, which tend to have lower entropy of vaporization, for example, Helium. The second exception is components which are associated in the liquid or vapour phase, for example, acetic acid.

Consider the majority of compounds that conform to Trouton's rule. The entropy can be described in terms of its translational, vibrational and rotational contributions. However, recognizing that for most molecules the vibrational and rotational part of the entropy is the same for the liquid and gas phases, the entropy change for these contributions is now zero. Thus, the only contribution of molecular motion that needs to be considered to account for Trouton's rule is changes in translation. Generally, translation can be interpreted as the external movement of the molecule. It is confined to the molar volume minus the volume occupied by the molecule itself. In the liquid phase, the available volume (V^{L}_{trans}) is usually less than a few percent of the total liquid volume. The translational part of the entropy of vaporization can be calculated via

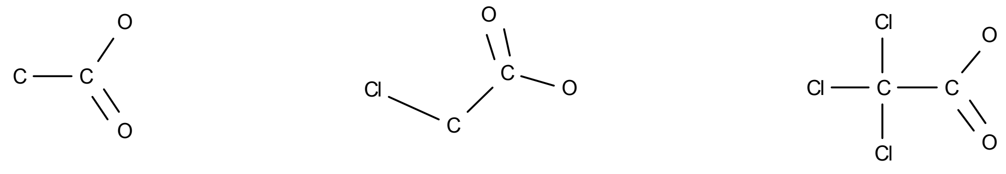
$$\Delta S^{vap}_{trans} = R\ln\left(V^{V}_{trans}/V^{L}_{trans}\right) \qquad (3\text{-}6)$$

It is interesting to note that the entropy of vaporization is insensitive to the ratio V^{V}/V^{L}. However, V^{L}_{trans} should decrease with increasing attractive forces. Thus the effect of volume change would be only on the enthalpy of vaporization.

3.3 Inductive and Resonance Effect

The terms "induction" and "resonance" refer to the electronic effects that atoms or functional groups may have within a compound. These effects depend on the valence, electronegativity of atoms, bonding order and molecular geometry of a molecule or functional group.

In strict definition, the inductive effect is an experimentally observable effect on the transmission of charge through a chain of atoms by electrostatic induction. A more simple definition is the withdrawal or donation of electrons through sigma bonds such as saturated (sp^3) carbon atoms. The inductive effect of an atom or functional group is a function of that group's electronegativity, bonding order, charge and position within a structure. Atoms or functional groups that are electronegative relative to hydrogen such as OH, F, Cl, etc, have a negative inductive effect (-I) or are polarized partially negative, depending on their bonding order. Thus these atoms withdraw electron density through the single bond structure of a compound. Consider the case of acetic acid, chloroacetic acid and trichloroacetic acid (Figure 3-1). All these structures can ionise (loss of proton from carboxyl OH). The only difference between these structures is the degree of chlorine substitution. Chlorine is electronegative and thus is polarised partially negative. Thus, they stabilize a negative charge and enhance the ionisation of an acid. They also induce a dipole moment, discussed later. Consider the pKa differences between acetic acid and chloroacetic acid. Furthermore, the more chlorine atoms (or electronegative atoms) present, the greater the total inductive effect and the ease of ionisation (lower pKa). Consequently, the electronic effect in this example is being induced through single saturated (sp^3) carbon atoms. Atoms of functional groups that are electron donating (hydrocarbons, anions) have a positive inductive effect (+I) or polarised partially positive. These groups can stabilise positive charges, for example, in reactions such as protonation of bases.



Acetic acid	Chloroacetic acid	Trichloroacetic acid
pKa = 4.76	pKa = 2.87	pKa = 0.64

Figure 3-1: pKa's of acetic acid, chloroacetic acid and trichloroacetic acid illustrating inductive effect (Hart et al (1995)).

Resonance may be defined as the bonding or sharing of electrons between more than two atoms. The classical example of resonance is provided by the pi-bonding system of benzene. Benzene is a six membered ring composed of six sp^2 hybridised carbon atoms in a plane and sharing 6 pi electrons. It can be represented by a kekule (Figure 3-2) structure which suggests an alternating single and double bond bonding pattern. This representation does not reflect the true electronic structure of benzene since all 6 pi electrons are shared equally by of the six carbon atoms. Thus the inscribed circular structure is more adequate considering the compound is now saturated. Consider for example, that the normal boiling point of cyclohexane (353.9 K) is lower than cyclohexene (356.1 K), with the latter component having a single double bond. However, benzene (353.3 K) has a similar in measure normal boiling point as cyclohexane. The effect of resonance is that now benzene has a greater stability, quite similar to that of cyclohexane.

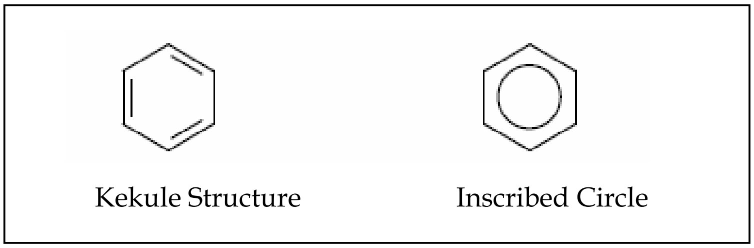


Figure 3-2: Kekule and inscribed circle structure of benzene.

Thus for resonance phenomena to exist, a ‘conjugated’ electronic structure must be present and the atoms involved in this system must be coplanar or adopt a coplanar conformation. This type of resonance effect exerted by an atom or functional group is determined by the electronic nature of the group. Each of these characteristics or requirements of resonance are described more in detail in common organic textbooks.

As described above, induction involves the electronic effects of atoms and functional groups through saturated carbons. Resonance involves the sharing or delocalization of electron pairs over more than two atoms and requires conjugation and coplanarity. To compare and contrast these two electronic effects on group contribution, consider for example, the electronic effects of an alcohol group (OH). This group is a withdrawer by induction (-I) and an electron donor by resonance. When placed in a structure where its resonance effects are ‘insulated’ by single bonds, only its electron withdrawing inductive effect will apply. When positioned within a structure where it can participate in delocalization of pi electrons, it will function as a strong electron donor. In both scenarios the polarizability of the alcohol functional group is different, because of the group's relative position with its neighbouring atoms. The polarizability of an atom has a major influence on the dipole moment and its relation to the normal boiling point will be discussed later in this chapter. Since the Cordes and Rarey method defined their groups according to its neighbouring atoms; this is the reason why the method produced the lowest average absolute deviation for almost all functional groups among all group contribution methods.

3.4 Intermolecular Forces

3.4.1 London Forces

London forces are weak attractive forces that are important over only extremely short distances. They exist for all types of molecules in condensed phases but are weak for small molecules. London forces are the only kind of intermolecular forces present among symmetrical non-polar molecules. Without London forces, these molecules could not condense to form liquids or solidify to form solids. Although van der Waal forces generally refer to all intermolecular attractions, it is also used interchangeably with London forces.

London forces result from the attraction of the positively charged nucleus of one atom to the electron cloud of an atom of another molecule. This induces temporary dipoles in neighbouring atoms or molecules. As electron clouds become larger and more diffuse, they are attracted less strongly by their own nuclei. Thus, they are more polarized by adjacent nuclei. Polarizability increases with increasing sizes of molecules and therefore with increasing numbers of electrons. Therefore, London forces are generally stronger for molecules that are larger or have more electrons. The increasing effectiveness of London forces of attractions occurs even in the case of some polar covalent molecules. For example, it accounts for the increase in boiling point in the sequences $HCl < HBr < HI$ and $H_2S < H_2Se < H_2Te$, which involve nonhydrogen-bonded polar covalent molecules. The difference in electronegativities decrease in these sequences, and the increasing London forces override the decreasing permanent dipole-dipole forces. An example of London forces on a homologous series is shown later.

3.4.2 Dipole moment

The dipole moment of a bond is defined as the product of the total amount of positive or negative charge and the distance between their centroids. In a molecule with only one covalent bond, the dipole moment of the whole molecule is identical to the dipole moment of the bond. Molecules with dipole moments are attracted to one other because they align themselves in such a way that the positive end of one dipole is close to the negative end of another dipole. These electrostatic attractive forces, called dipole-dipole interactions, are stronger than van der Vaals forces but not as strong as ionic or covalent bonds.

To describe the effect of the dipole moment, consider boiling points of ethers and alkanes of comparable molecular weight. Ethers generally have a higher boiling point than alkanes because both the Van der Vaals forces and dipole-dipole interactions in ether, must be overcome for ethers to boil (Table 3-1). The resultant increase of the attractive forces between the molecules, which results from the dipole-dipole interactions, increases the enthalpy of vaporization. Also, due to the greater molecular interaction, the disorder of the system increases, however nature tends to keep the entropy change small, thus increasing the boiling point. Therefore, with higher molecular interaction, the boiling point must increase, since ΔS^V is also weakly dependant on temperature.

Table 3-1: Comparative boiling points of alkanes, ethers, alcohols and amines (Atkins (1994)).

Compound	**NBP**	**Compound**	**NBP**	**Compound**	**NBP**
$CH_3CH_2CH_3$	-42.1	$CH_3CH_2CH_2CH_3$	-0.5	CH_3 $CH_2CH_2CH_2CH_3$	36.1
CH_3OCH_3	-23.7	$CH_3OCH_2CH_3$	10.8	$CH_3CH_2OCH_2CH_3$	34.5
CH_3CH_2OH	78.0	$CH_3CH_2CH_2OH$	97.4	$CH_3CH_2CH_2CH_2OH$	117.3
$CH_3CH_2NH_2$	16.6	$CH_3CH_2CH_2NH_2$	47.8	$CH_3CH_2CH_2CH_2NH_2$	77.8

(NBP – Normal boiling point (°C))

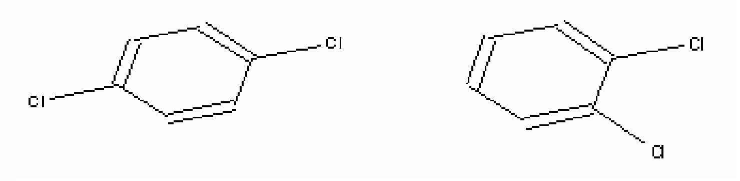
Table 3-2 presents dipole moments for some commonly encountered bonds. It can be seen that the effect of the neighbour of a group is a significant factor in the boiling point prediction. Consider the group -CH_2, the dipole moment of C-C bond and C-O bond (RCH_2-CH_3 and RCH_2-OH in Table 3-1, respectively) is significantly different. If the group contribution method is not able to distinguish between the groups, the regression procedure finds a parameter value for the -CH_2 group between the two components, which leads to higher outliers. Consequently, group contribution needs to incorporate this behaviour, but not as Marrero and Pardillo (1999) did by including every bond. The idea of Cordes and Rarey (2002) defining the neighbourhood of the groups would be more scientific.

Table 3-2: Dipole moments of some commonly encountered bonds (Atkins (1994)).

Bond	Dipole moment (D)	Bond	Dipole moment (D)
C-C	0	C-Cl	1.5
C-H	0.4	C-Br	1.4
C-N	0.2	C-I	1.2
C-O	0.7	H-N	1.3
C-F	1.4	H-O	1.5

The definition of a structural group which includes the neighbourhood of the group would be sufficient to estimate the normal boiling point of molecules where there is a single covalent bond. However, it becomes more complicated for molecules that contain more than one covalent bond. The geometry of the molecule and therefore the vector sum of all individual bond dipole moments has to be taken into account when determining the overall dipole moment of the molecule. The vector sum takes into account both the magnitudes and the direction of the bond dipoles.

Consider the dipole moments for p-dichlorobenzene and o-dichlorobenzene (Figure 3-3). Both components are isomers, of which the difference in dipole moments is 2.5 D. This results in a boiling point difference of about 6K. Group contribution cannot distinguish between these molecules. The effect of the dipole moment becomes even greater with more electronegative groups, for example, CN, OH, NH_2 etc.



p-dichlorobenzene – 0 D o-dichlorobenzene – 2.5 D

Figure 3-3: Dipole moments for p-dichlorobenzene and o-dichlorobenzene (DDB).

3.4.3 Intermolecular Hydrogen bonding

Hydrogen bonded to oxygen, nitrogen or fluorine can form a weak association with a second oxygen, nitrogen or fluorine of a different molecule (Figure 3-4). This association is known as intermolecular hydrogen bonding. The length of the covalent bond between an oxygen and hydrogen within an alcohol group is 0.96 Ǻ. The hydrogen bond between an oxygen of one molecule and hydrogen of another molecule is twice as long (1.69 – 1.79 Ǻ). So the hydrogen bond is not as strong as an oxygen-hydrogen covalent bond in the alcohol group, but is stronger than some dipole-dipole interactions.

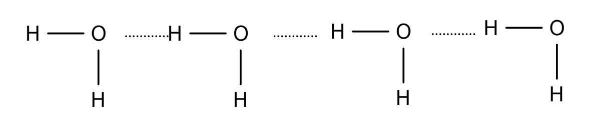


Figure 3-4: Hydrogen bonding in water

Thus, the increase in attractive forces in the liquid phase increases the heat of vaporization, as explained earlier. The extra energy required to break these bonds is the main reason why molecules with hydrogen bonds have much higher boiling points. The boiling point of water illustrates this behaviour, and a boiling point of 100 °C, with a molecular weight of 18. The closest alkane in size is methane, with a molecular weight of 16, which has a normal boiling point of -167.7 °C. Also, alcohols and amines, molecules with hydrogen bonding, generally have higher boiling points (Table 3-1) than alkanes and ethers of comparable molecular weight.

The strongest hydrogen bonds are linear, where two electronegative atoms and the hydrogen between them, lie in a straight line. Nitrogen is less electronegative than oxygen, which means hydrogen bonds between amines are weaker than hydrogen bonds in alcohols. Amines, therefore, have lower boiling points than alcohols (Table 3-1) of comparable molecular weight.

3.4.4 The Potential Energy of Interaction

The properties observed for organic compounds on the macroscopic level are determined by the properties of individual molecules and the interactions between them. The polar or non-polar character of a molecule will clearly be important in determining the nature of its interactions with other molecules. These interactions can be considered the result of the effects described above. Thus, thermodynamic properties of any pure substance can be determined by these forces which operate between the molecules. Thus, when considering molecules with similar groups but a different nature, these effects cannot be differentiated entirely, within the scope of group contribution estimation.

Molecules have kinetic energy as a result of their velocities relative to some fixed frame of reference. They also have potential energy from their positions relative to one another. Molecules in the condensed phase are in a region of low potential energy due to the attractive forces exerted by the neighbouring molecules. By supplying energy in the form of heat, molecules in the liquid phase can acquire sufficient kinetic energy to overcome the potential energy of attraction and escape into the vapour phase. The vapour pressure will thus provide a means to measure the tendency of a molecule in a condensed phase to escape into the vapour phase. The larger the vapour pressure, the greater the escaping tendency. Thus, the observation of a large vapour pressure at a low temperature implies that relatively little kinetic energy is required to overcome the potential interactions between the molecules in the condensed phase.

The potential energy of interaction between molecules resulting from intermolecular forces needs to be overcome for the boiling point to be reached. These intermolecular forces (as described above) are the general reason for differences in boiling points occurring between compounds of comparable molecular weight. Table 3-3 presents typical potential energies for these interactions.

The ion-ion interaction has by far the highest potential energy. These types of compounds, generally referred to as ionic liquids, have no measurable vapour pressures and will not be considered in this work. Thus, molecules with hydrogen bonding tend to have higher boiling points than molecules with dipole-dipole interactions and London forces. This can be attributed to the higher potential energy needed to break this bond. Although the London potential energy is almost the same as the dipole-dipole interaction, these forces are only effective over short distances. Thus, it can be considered the weakest intermolecular force.

Table 3-3: Typical potential energies of charges and dipoles (Atkins (1994))

Interaction Type	Distance Dependence	Typical Energy (kJ mol^{-1})	Comments
Ion-Ion	1/r	250	Only between ions
Hydrogen Bond A-H…B	-	20	A,B = N, O or Fl
Ion-Dipole	1/r^2	15	
Dipole-Dipole	1/r^3	2	Between stationary polar molecules
	1/r^6	0.3	Between rotating polar molecules
London	1/r^6	2	Between all types of molecules

3.5 Intra-molecular Hydrogen bonding

Hydrogen bonding within the molecule leads to strong intra-molecular interaction resulting in a significant boiling temperature elevation. In cases, where steric effects force an intra-molecular hydrogen bond, the boiling temperature is much lower than expected. Such cases are (Figure 3-5) (a) o-nitrophenols, (b) o-hydroxy carbonyl aromats (for example, salicylic acid), (c) o-alkoxy benzaldehydes, (d) b-keto esters and (e) 1,3-dicarbonyl compounds. These are typical cases which exhibit intra-molecular hydrogen bonding. The normal boiling points for most of these types of components are generally not available. Consequently, a group contribution prediction should be considered as a very rough estimate.

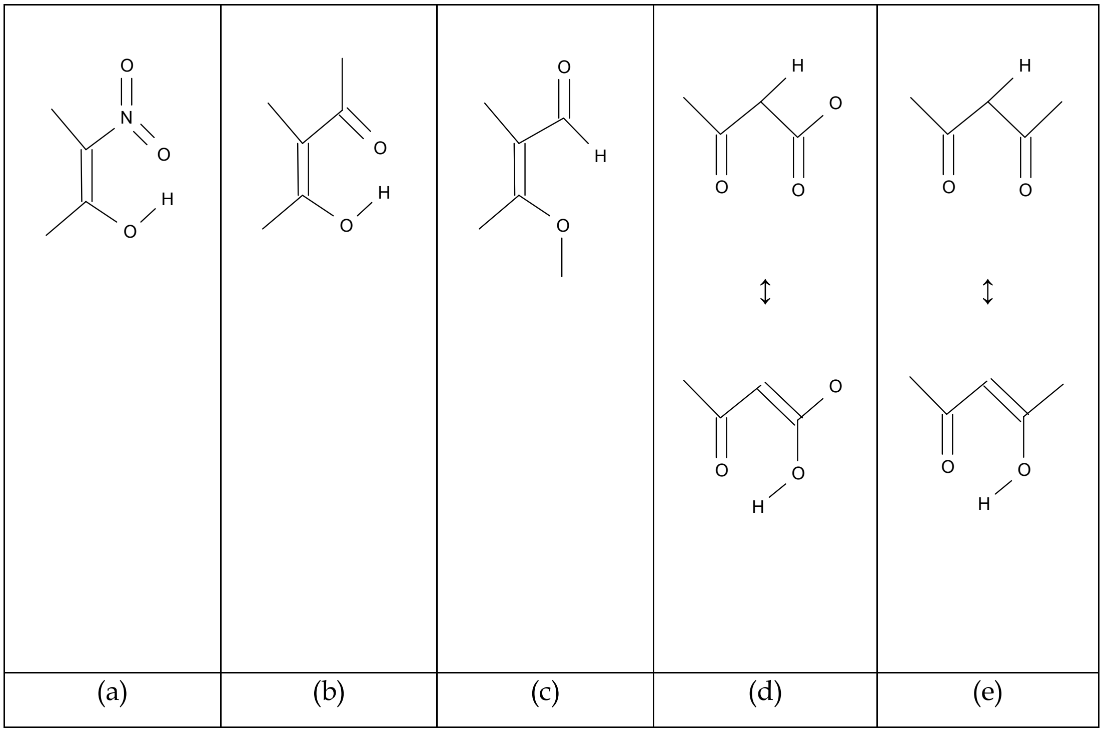


Figure 3-5: Intra-molecular hydrogen bonding

3.6 Acid/Base Interactions

There are numerous possible ways to define an acid and a base. One common way is the Lewis acid and base model:

- An acid is an electron pair acceptor.
- A base is an electron pair donor.

This definition is broader than the Bronsted-Lowery definition. In both definitions, H^+ is an acid and OH^- is a base, since for a proton to bind a base, it must accept a pair of electrons:

$H^+ (aq) + :OH^- (aq) \rightarrow H_2O (l)$

However, the Lewis definition extends beyond just the proton. For example, many metals ions can act as Lewis acids when they form complex ions:

$Fe^{3+} (aq) + 6CN^- (aq) \rightarrow Fe(CN)_6^{-3} (aq)$

The electrons which form the bond between the iron and cyanide ion start as lone pairs on the cyanide. The iron accepts the electrons and is thus an electron pair acceptor, Lewis acid; the cyanide donate the electrons and thus is an electron pair donor, Lewis base. Note that there are no protons in the reaction, but it is still an acid/base reaction.

The effect of the acid/base interactions is similar to the hydrogen bonding effect. There is a resultant increase in attractive forces between groups that can act as Lewis acids, bases or both. To consider the acid/base interactions, consider molecules where hydrogen bonding does not occur, for example thiol molecules (SH). Thiol molecules are amphoteric, i.e. can act as acid or base. The boiling point of ethanethiol, 36.3 °C with a molecular weight of 58, is higher than butane, -0.5 °C with a molecular weight of 62. However, 1,2-ethanedithiol has a considerably larger boiling point of 148.9 °C, with a molecular of 94 (Figure 3-6), than hexane, 69 °C with a molecular weight of 86. The resultant increase in boiling point is due to the acid/base interaction and the dipole-dipole interactions, of which the former has been discussed previously.

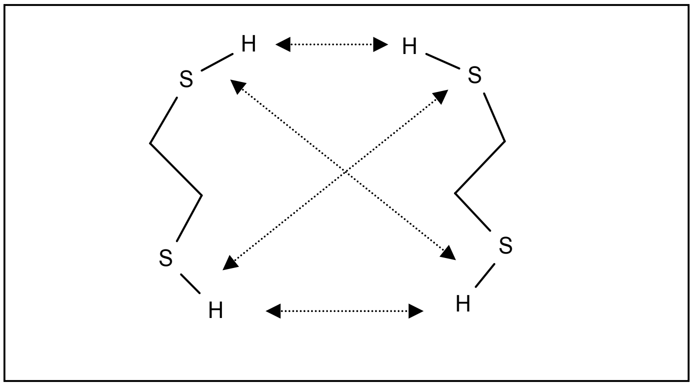


Figure 3-6: Acid/base effect on 1,2-ethanedithiol

3.7 Molecular size

The molecular size of a molecule can be interpreted as molecular volume, molecular surface area or molecular weight. In general, the boiling point increases with increasing molecular size. Consider for example, a n-alkane homologous series. A homologous series is a series of compounds in which each member differs from the next by a specific number and kind of atoms, for the case of a n-alkane homologous series, it differs by a $–CH_2$-group. The volume of the molecule increases linearly with each $–CH_2$ group added. Generally, molecules tend to adhere to a more or less spherical form at which their outer surface area should approximately increase with $n_{CH2}^{2/3}$, once a certain length is reached.

The saturated hydrocarbons of n-alkanes are non-polar molecules. Thus, the only significant intermolecular forces are London forces, which were discussed earlier. The trends as depicted in Figure 3-7 are due to the increase in effectiveness of the London forces. Figure 3-7 shows normal boiling for n-alkanes as a function of number of atoms together with a correlation (regression performed over hydrocarbons) using the expression:

$$T_b = -409.13+474.50*n^{0.2572} \qquad (3\text{-}11)$$

In case of small chains (1 to 4 CH_2-groups), ΔH^{vap} increases linearly with the number of CH_2-groups. Thus, in general, the estimation of the normal boiling temperature of the first few components of a functional series is more than expected. For very large molecules, a mutual contact of the complete outer surface becomes more difficult (increasing free volume) and the increase of ΔH^{vap} is less than estimated.

This is not, however, always the case. Consider for example alkanes, alkenes (only one double bond) and alkynes (only one triple bond). Alkane compounds commonly have a higher boiling point than alkenes (Figure 3-8), which could be attributed to the larger molecular size of alkanes. However, alkyne compounds have a higher boiling point than both alkane and alkene compounds, even though the molecular size of alkyne compounds is smaller.

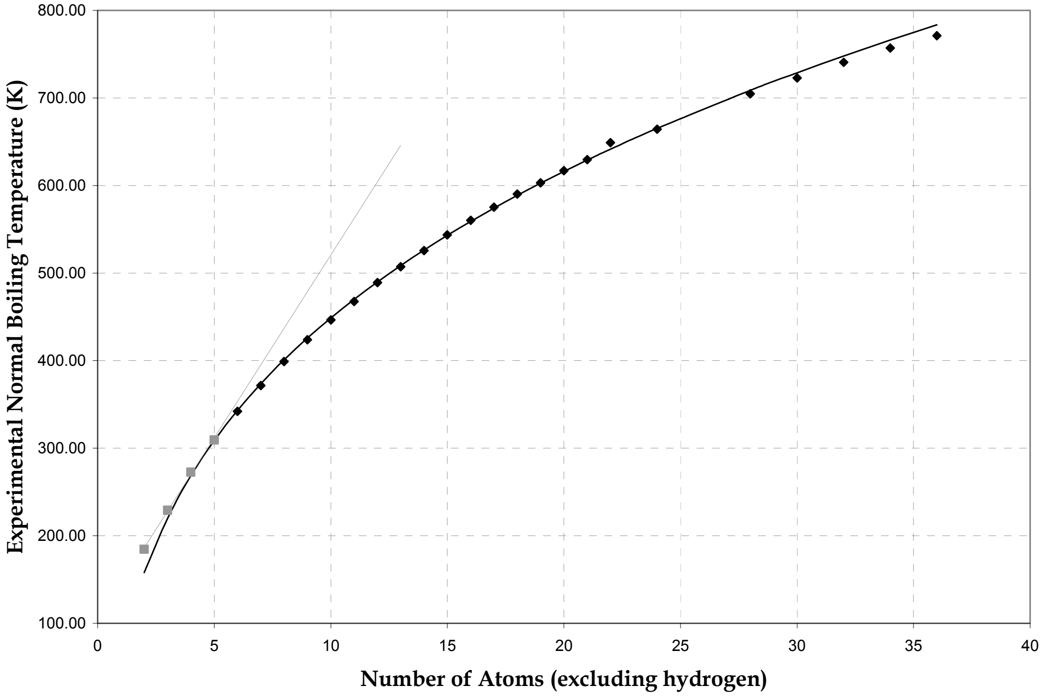


Figure 3-7: Normal boiling temperatures of n-alkanes as a function of number of atoms (DDB).

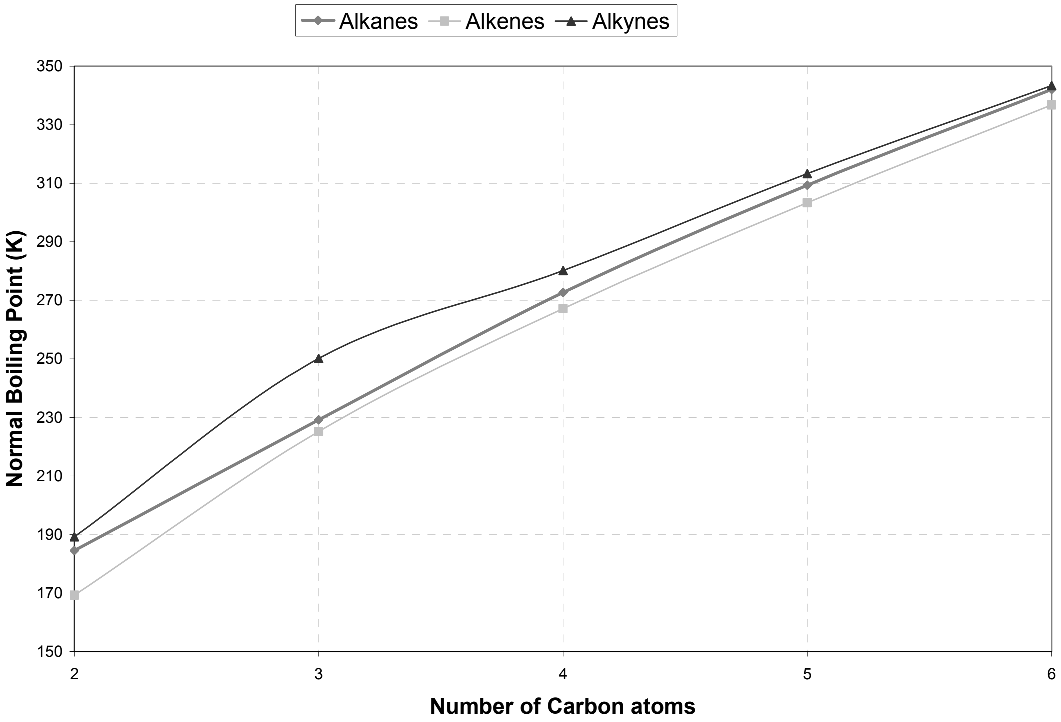


Figure 3-8: Normal boiling points for series range (C_2 – C_6) of alkanes, alkenes and alkynes (DDB).

The actual difference in boiling point is a result of the change in dipole moment, discussed earlier. Consider the ethane series, ethene has a single double bond and a higher dipole moment than ethane. However, the entropy change (Table 3-3) is extremely small such that, the resultant decrease in molecular size is more influential than the increase in attractive forces, which comes about from the polar double bond. This can be seen by the smaller enthalpy of vaporization for ethene as compared to ethane. However, for acetylene, now with one triple bond, the entropy change is significantly larger, which illustrates greater disorder in the system. Thus the molecular size is now less significant with the enthalpy of vaporization being greater than that of ethane.

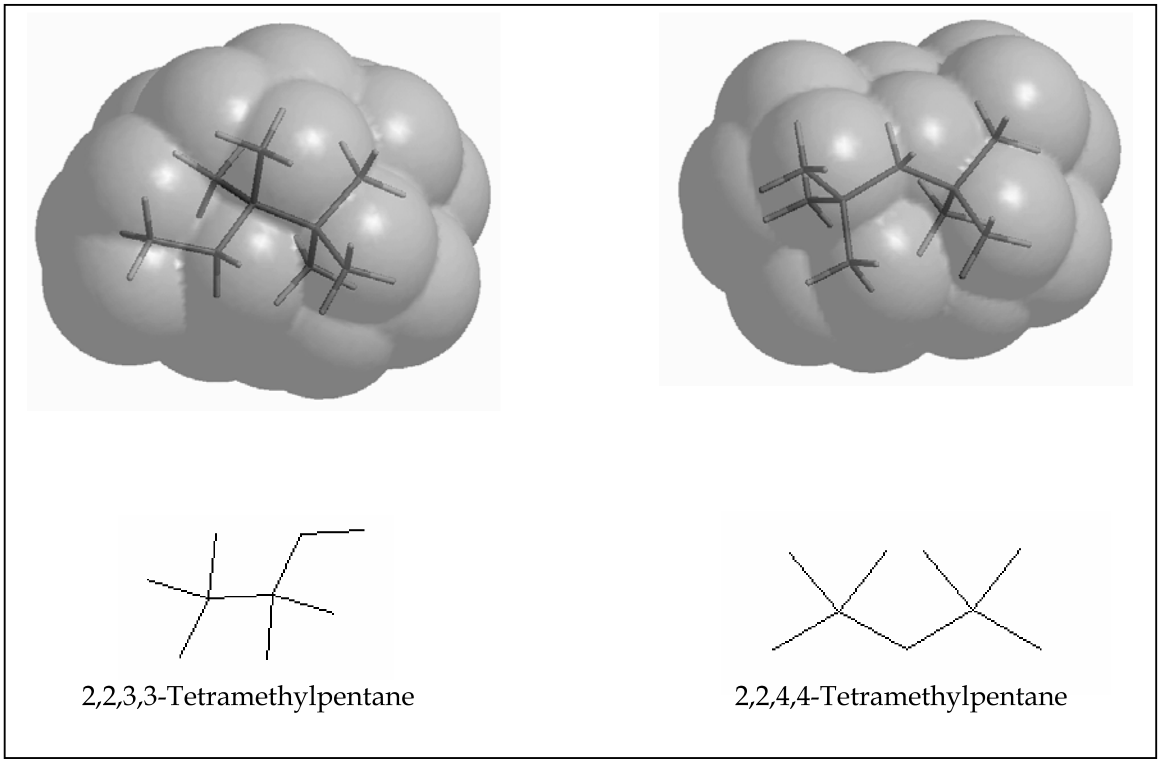
Table 3-4: Normal boiling point, enthalpy of vaporization and entropy of vaporization of ethane, ethene and acetylene at atmospheric conditions (DDB).

	NBP (K)	**ΔH^{vap} (J.mol^{-1})**	**ΔS^{vap} (J.mol^{-1}.K^{-1})**
Ethane	184.49	14681.97	79.5814
Ethylene	169.25	13511.85	79.83369
Acetylene	189.15	16659.27	88.07437

3.8 Steric Hindrance

Steric hindrance (or steric strain) is the strain put on the molecule when atoms or groups are too close too each other, which causes repulsion between the electron clouds of the atoms or groups. In general, it is considered that the increasing number of branches on a hydrocarbon chain decreases the molecular area of the molecule. The molecule now becomes more compact, nearly spherical in shape, thus decreasing the boiling point. This differentiation results in different normal boiling points of isomer compounds. However, a sterically hindered compound is considered less stable than that of its isomers. Thus, the potential energy of the hindered molecule is higher then that of its conformers, which would mean that extra energy is needed to overcome this strain, or potential energy. This would imply a greater boiling point for the hindered molecule.

Consider two isomers of nonane, 2,2,3,3-Tetramethylpentane (1) and 2,2,4,4-Tetramethylpentane (2) (Figure 3-9). Both molecules have the same number of branches; however, component (1) has the branches one bond apart and component (2) two bonds apart. Thus, theoretically, with component (1) there is a greater strain on the bond with four branches, and the boiling point should be greater than that of component (2). To verify this statement, Table 3-4 has the normal boiling points, the Connolly molecular area[4] and Connolly molecular 'solvent excluded' volume[5] of the above two components (Connolly (1996)). The molecular area and molecular volume were calculated using a simulation package (ChemOffice 6.0). As expected the normal boiling point is higher for component (1), even though the molecular area and molecular volume is smaller.



2,2,3,3-Tetramethylpentane

2,2,4,4-Tetramethylpentane

Figure 3-9: Molecular structures for 2,2,3,3-tetramethylpentane and 2,2,4,4-tetramethylpentane

[4] The contact surface created when a spherical probe sphere (representing the solvent) is rolled over the molecular model.

[5] The volume contained within the contact molecular surface.

Table 3-5: Normal boiling point, molecular area and molecular volume to illustrate steric hindrance of alkanes (DDB and ChemOffice)

	NBP (K)	**Molecular Area (Ǻ²)**	**Molecular Volume (Ǻ³)**
(1) 2,2,3,3-Tetramethylpentane	413.4	168.1	163.7
(2) 2,2,4,4-Tetramethylpentane	395.5	171.7	165.4
(3) Decane	446.5	217.0	180.4
(4) 3,3,5-Trimethylheptane	428.7	192.7	181.7
(5) 2,2,3,3-Tetramethylhexane	433.0	185.6	181.0
(6) 2,2,4,5-Tetramethylhexane	421.4	190.9	182.7
(7) 2,2,3,4-Tetramethylhexane	427.5	187.7	181.2

Chapter Four

Mathematical and Software Considerations

4.1 Development of Regression Algorithm

Since model development for the group contribution method is of key importance to the method, the quest is to develop a regression algorithm to process all models efficiently. Model development entails the development of a relationship (Equation 4-1) of the normal boiling point of a compound as a function of structurally dependant parameters (group contribution parameters). These parameters are determined by summing the number frequency of the compound multiplied by its contribution.

$$f(T_b) = \Sigma N_i C_i \qquad (4\text{-}1)$$

For the simultaneous regression of the model and group parameters, a special algorithm was developed consisting of an inner and outer regression loop. The outer loop optimizes for the non-linear constants in order to minimize the sum of squares between the calculated and experimental normal boiling point temperatures. This common objective function leads to slightly higher mean deviations; however, it decreases the average deviations of outliers which exhibit high deviations. The inner loop performs a multi-linear least squares fit on the linear group parameters.

4.1.1 Non-linear Regression

The criteria for choosing a non-linear algorithm are based upon the efficient use of the algorithm to handle the different types of non-linear functions. Since model testing forms an integral part of the method, many different forms of mathematical functions will be sought to obtain the best model for the relationship between the normal boiling point and the group contribution parameters. The input or required form for the non-linear algorithm must also be taken into consideration. If an algorithm requires the input form to be derived from the non-linear equation, then this will be tedious work for the many different types of models that will be tested.

The Simplex method (Nelder and Mead (1965)) satisfies the above criterion. The method requires only function evaluations; not derivatives. However, it is not very efficient in terms of the number of function evaluations that it requires. Also, the method is generally slow in all likely applications, but, considering the number of non-linear parameters to be optimized is in the range of 0 to 6, it would be sufficiently quick and more importantly stable to perform the regression. The simplex method may frequently also be the best method to use, on a problem where the computational burden is small.

4.1.1.1 Description of the Simplex Method

A simplex is a geometrical figure, in n-dimension, consisting of (n + 1) points and all their interconnecting line segments, polygonal faces, etc. In two dimensions, the simplex is an equilateral triangle, in three dimensions it a tetrahedron. x_0, x_1, … , x_n are the (n + 1) points in the n-dimensional space defining the current simplex and y_i are the function values at $x_{i.}$ The suffixes *h* and *l* are defined as "high" and "low" respectively, as denoted in Equations 4-2 and 4-3:

$$y_h = \max_i(y_i) \tag{4-2}$$

$$y_l = \min_i(y_i) \tag{4-3}$$

x_m is further defined as the centroid of the points with $i \neq h$, and define [x_i x_j] as the distance between x_i and x_j. For each stage in the process x_h is replaced by a new point, this is done using three operations viz reflection, contraction and expansion.

4.1.1.2 Reflection

The reflection (Figure 4-1) of x_s is denoted by x_r, and its co-ordinates are defined by the Equation 4-4:

$$x_r = (1 + \alpha)\, x_m - \alpha x_s \qquad (4\text{-}4)$$

where α is the reflection coefficient, a positive constant. Thus x_r is on the line joining x_s and x_m, on the far side of x_m from x_s with $[x_r\ x_m] = \alpha[x_s\ x_m]$. If y_r lies between y_h and y_l, then x_s is replaced by x_r and the procedure starts again with a new simplex.

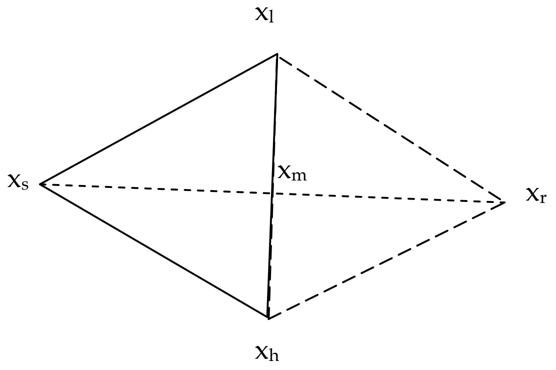


Figure 4-1: Simplex for reflection

4.1.1.3 Expansion

If $y_r < y_l$, i.e. the reflections has produced a new minimum, then x_r is expanded (Figure 4-4) to x_e by the equation

$$x_e = \gamma x_r + (1 - \gamma) x_m \qquad (4\text{-}5)$$

where γ is the expansion coefficient, which is greater than unity. It is the ratio of the distance $[x_e\ x_m]$ to $[x_r\ x_m]$. If $y_e < y_l$, then x_s is replaced by x_e and the process is re-started. But if $y_e > y_l$, then the expansion has failed, and x_s is replaced by x_r before re-starting.

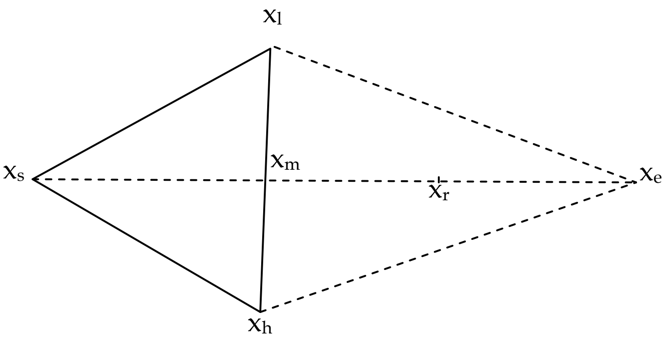


Figure 4-2: Simplex for expansion

4.1.1.4 Contraction

On reflecting x_s to x_r, if $y_r > y_h$ i.e. by replacing x_s by x_r leaves y_r the maximum, then a new x_s is defined to be either the old x_s or x_r, whichever has the lower y value, and form the equation

$$x_c = \beta x_s + (1 - \beta)x_m \qquad (4\text{-}6)$$

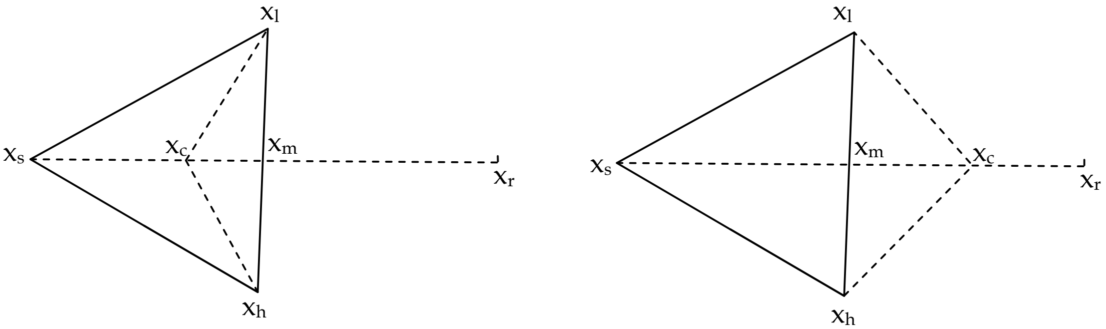


Figure 4-3: Simplex for contraction

The contraction coefficient β, lies between 0 and 1 and is the ratio of the distance [x_c x_m] to [x_s x_m]. x_c then replaces x_s and the process is re-started, unless $y_c > \min(y_h, y_r)$, i.e. the contracted point (Figure 4-3) is worse than the better of x_h and x_r. For such a failed contraction, all the x_i are replaced by $(x_i + x_l) / 2$ and the process is restarted.

The flow diagram of the simplex algorithm is presented in Figure 4-4.

4.1.2 Multi-Linear Regression

Function minimization of a set of linear equations can be performed quickly and effectively by using a multi-linear least squares regression. The general form of this kind of model is:

$$y_i = a_0 + \sum_{j=1}^{M} a_j x_{ij} \tag{4-7}$$

where x_{i1}, x_{i2}, … , x_{iM} are arbitrary fixed functions of x.

A least squares solution to the above model can be found in many mathematical textbooks. However, on comparing this model to the proposed model (Equation 4-1) the above model contains a constant a_0. Subsequently, it is not possible to perform a regression of contributions of individual groups separately, as this would lead to different and incompatible values for the constant a_0. This was the reason why in the previous methods, the regression was performed on the full set of data and this made it difficult to identify unreliable data, inappropriate group definitions, etc.

The model for the proposed method for the least squares fit is presented in Equation 4-8, where M is the number of structural groups, including second-order corrections.

$$y_i = \sum_{j=1}^{M} a_j x_{ij} \tag{4-8}$$

Since it was not successful to find the equations for a linear regression of this type of model in common mathematical textbooks, these equations had to be derived. The objective function (S) is defined as the sum of squares of the deviation between the calculated and experimental normal boiling points (Equation 4-9). The experimental normal boiling points are now defined as y_i and N is the total number of data points.

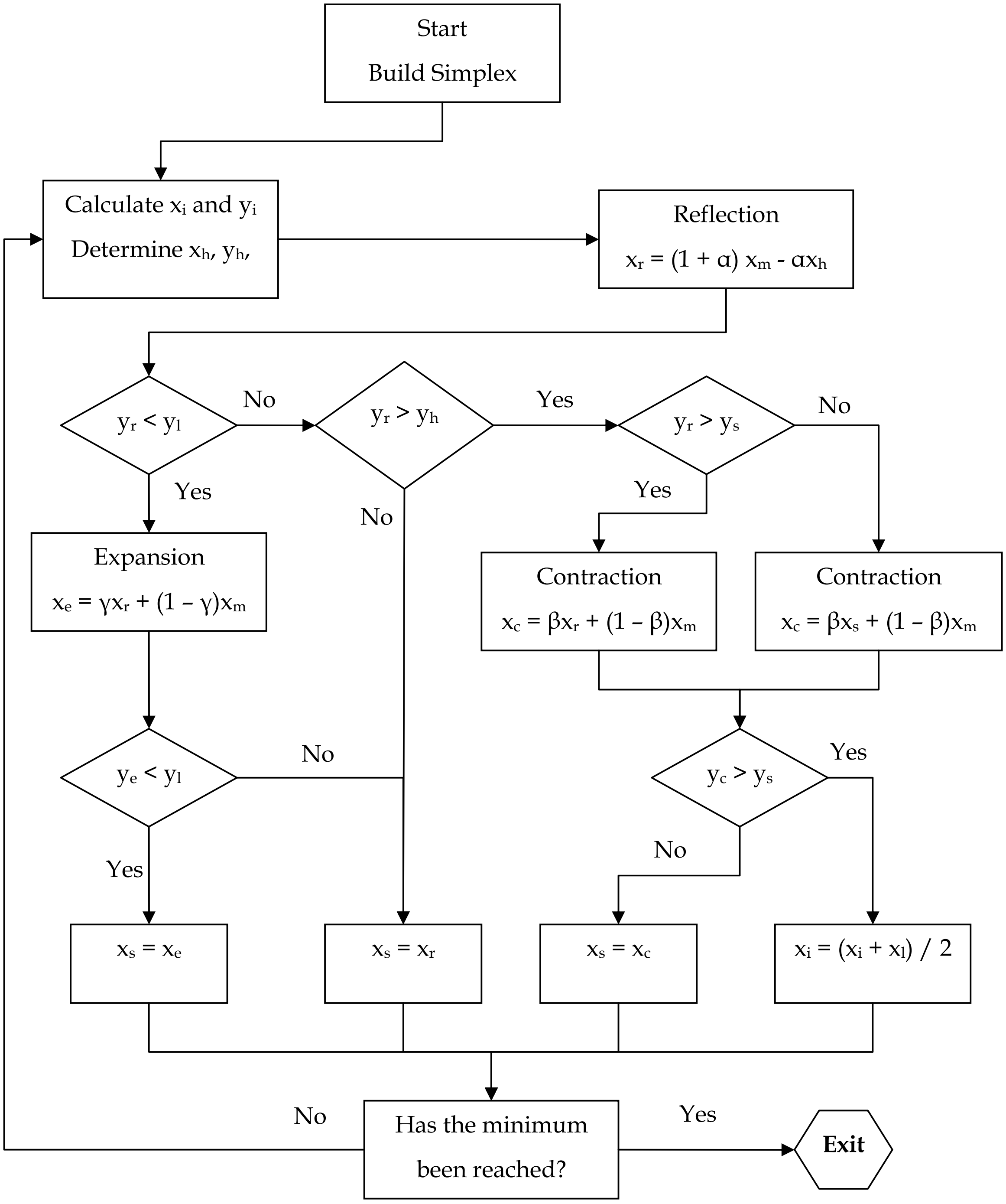

Start
Build Simplex
Calculate x_i and y_i
Determine x_h, y_h,
Reflection
$x_r = (1 + \alpha)\ x_m - \alpha x_h$
$y_r < y_l$
No
$y_r > y_h$
Yes
$y_r > y_s$
No
Yes
No
Yes
Expansion
$x_e = \gamma x_r + (1 - \gamma)x_m$
Contraction
$x_c = \beta x_r + (1 - \beta)x_m$
Contraction
$x_c = \beta x_s + (1 - \beta)x_m$
No
$y_e < y_l$
Yes
$y_c > y_s$
No
Yes
$x_s = x_e$
$x_s = x_r$
$x_s = x_c$
$x_i = (x_i + x_l)\ /\ 2$
No
Has the minimum been reached?
Yes
Exit

Figure 4-4: Flow Diagram of the Simplex Algorithm

$$S=\sum_{i=1}^{N}\left[y_i-\sum_{j=1}^{M}a_jx_{ij}\right]^2 \quad \text{or} \quad S=\sum_{i=1}^{N}\left[y_i^2-2y_i\sum_{j=1}^{M}a_jx_{ij}+\left(\sum_{j=1}^{M}a_jx_{ij}\right)^2\right] \tag{4-9}$$

At the minimum S, the partial derivatives of S with respect to the coefficients a_k are zero:

$$\frac{\partial S}{\partial a_k}=0=-2\sum_{i=1}^{N}y_ix_{ik}+2\sum_{j=1}^{M}a_j\sum_{i=1}^{N}x_{ik}x_{ij} \tag{4-10}$$

This leads to a set of M linear equations

$$\sum_{i=1}^{N}y_ix_{ik}=\sum_{j=1}^{M}a_j\sum_{i=1}^{N}x_{ik}x_{ij} \quad for\ k=1\ to\ M \tag{4-11}$$

or

$$\begin{vmatrix} \sum_{i=1}^{N}x_{i1}x_{i1} & \cdots & \sum_{i=1}^{N}x_{i1}x_{iM} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ \sum_{i=1}^{N}x_{iM}x_{i1} & \cdots & \sum_{i=1}^{N}x_{iM}x_{iM} \end{vmatrix} * \begin{vmatrix} a_1 \\ \mathrm{M} \\ a_M \end{vmatrix} = \begin{vmatrix} \sum_{i=1}^{N}y_ix_{i1} \\ \mathrm{M} \\ \sum_{i=1}^{N}y_ix_{iM} \end{vmatrix} \tag{4-12}$$

which can easily be solved for a_1, a_2, … , a_M, which are the group contribution constants.

4.1.3 Overall Flow Diagram of the Regression Algorithm

The overall flow diagram for the regression algorithm is presented in Figure 4-5. The 'MAIN' block provides the input, starting values, step-width, etc., for the algorithm. The outer loop 'DSIM' then solves for the non-linear constants, which requires a function evaluation (y_i). This evaluation must be performed on a set of optimized linear constants. This is done by the inner loop 'LINREG' and 'SIMQ'. The 'AUX' block obtains the objective function and also prepares the function (y_i) for the least squares fit. The bypassing of the 'DSIM' block provides a regression of structural group constants. In this case, the previous optimized values for the non-linear constants are used.

4.2 Automatic Fragmentation Procedure

Fragmentation of molecules into defined structural groups can be a tedious and time consuming procedure. However, this is obviously required as a means to test the predictive capability of any method. As mentioned in Chapter 2, for the Cordes and Rarey (2002) method, the authors developed an automatic fragmentation procedure fragment molecules into its respective structural groups. The same procedure was used here. The input requirement is an ink file, which contains the structural definition of the groups.

4.2.1 Ink File

The ink file is basically a text file with the extension being .ink. It provides structural information for the defined groups. Thus for any method, an ink file can be developed and the automatic procedure will fragment the molecules according to the group's definition.

To describe the structural definition of a group in an ink file; it can be best explained using an example. Figure 4-6; a carboxylic acid, can be interpreted as: (the numbers in brackets at the beginning of each line are used for explanation purposes only, but is not defined in the ink file)

Line 1:
Contains the name and a shortened name of the group between the section sign operator (§). This shortened name is used as verification in the filter language; explained later on.

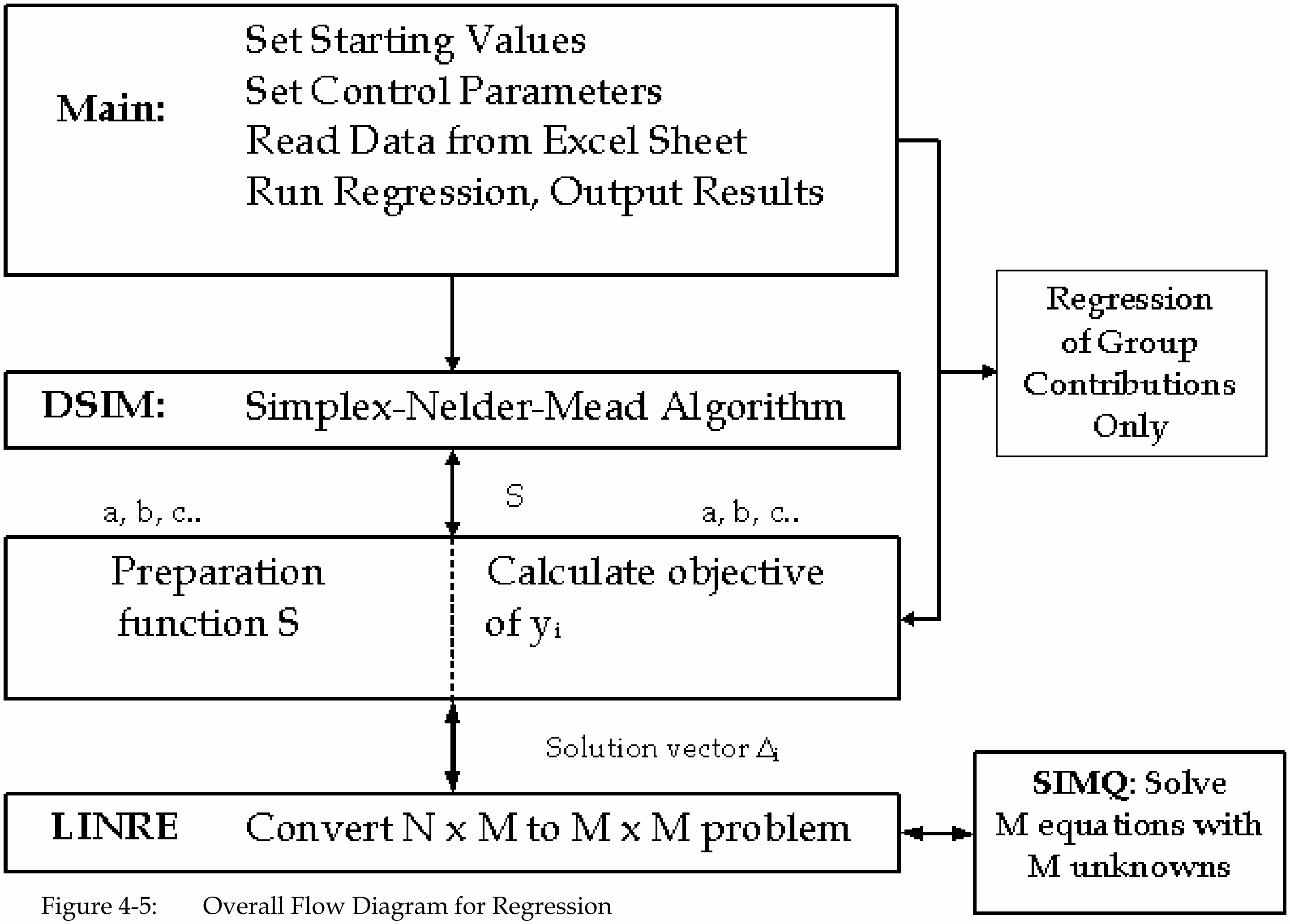

Main:
Set Starting Values
Set Control Parameters
Read Data from Excel Sheet
Run Regression, Output Results
Regression of Group Contributions Only
DSIM:
Simplex-Nelder-Mead Algorithm
a, b, c..
S
a, b, c..
Preparation function S
Calculate objective of y_i
Solution vector Δ_i
LINRE
Convert N x M to M x M problem
SIMQ: Solve M equations with M unknowns

Figure 4-5: Overall Flow Diagram for Regression

```
 (1)     Carboxylic acid§COOH§
(2)       4  3  44  44
(3)      C 3 2 K 0 Ja
(4)      O 1 1 K 0 Ja
(5)      O 1 1 K 0 Ja
(6)      C 4 1 * 0 Nein
(7)      1 2 2 K
(8)      1 3 1 K
(9)      1 4 1 K
```

Figure 4-6: Example of a group definition in an ink file

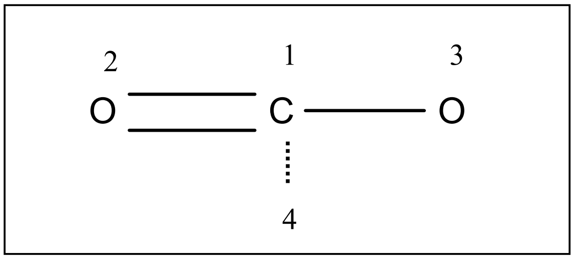


Figure 4-7: Carboxylic acid structure for the example above.

Line 2:

Incorporates the description of the structural group. The line has 4 items, each separated by a space. The first item is the number of atoms in the group, here '4'. The second item is the number of bonds, here '3'. Third and fourth item are main group and subgroup numbers, in this case it is identical, which is dependant on the method, for example, 2 separate numbers needed for UNIFAC.

Lines 3-6:

Information about the atom, which has 6 items. First item is element; second and third items are the maximum number of neighbours and minimum number of neighbours, respectively. For example, in Figure 4-7, the carbon atom can only have a maximum of three neighbours due to a double bond being present, and a minimum of two neighbours, as a result of the two oxygens ('2' and '3') being present. Fourth item is the neighbourhood of the atom, here for example, 'K' represents chain. Fifth item is the charge, '0' for no charge. Sixth item is whether the item should be included as part of the group definition. The term 'Ja' is used as verification to fragment the particular element as part of the group. The ‘Nein’ term is generally used to describe the neighbourhood of the group but is not fragmented as part of the group. This complies with the general rule that an element can only be fragmented once. For example, if atom ‘4’ is a –CH group, then the carboxylic acid requires this group for the fragmentation to occur, but the –CH is not fragmented as part of a carboxylic acid. Instead it will be fragmented separately. This definition is used for the fragmentation to capture the neighbourhood of the group.

Lines 7-9:
Information about the bond which has 4 items. First and second items are the atom reference numbers, for example, ‘1’ refers to line 3, ‘2’ refers to line 4, and so on. Third item is the bond type, ‘1’ – single bond, ‘2’ – double bond, ‘3’ – triple bond. Fourth item refers to the neighbourhood of the group.

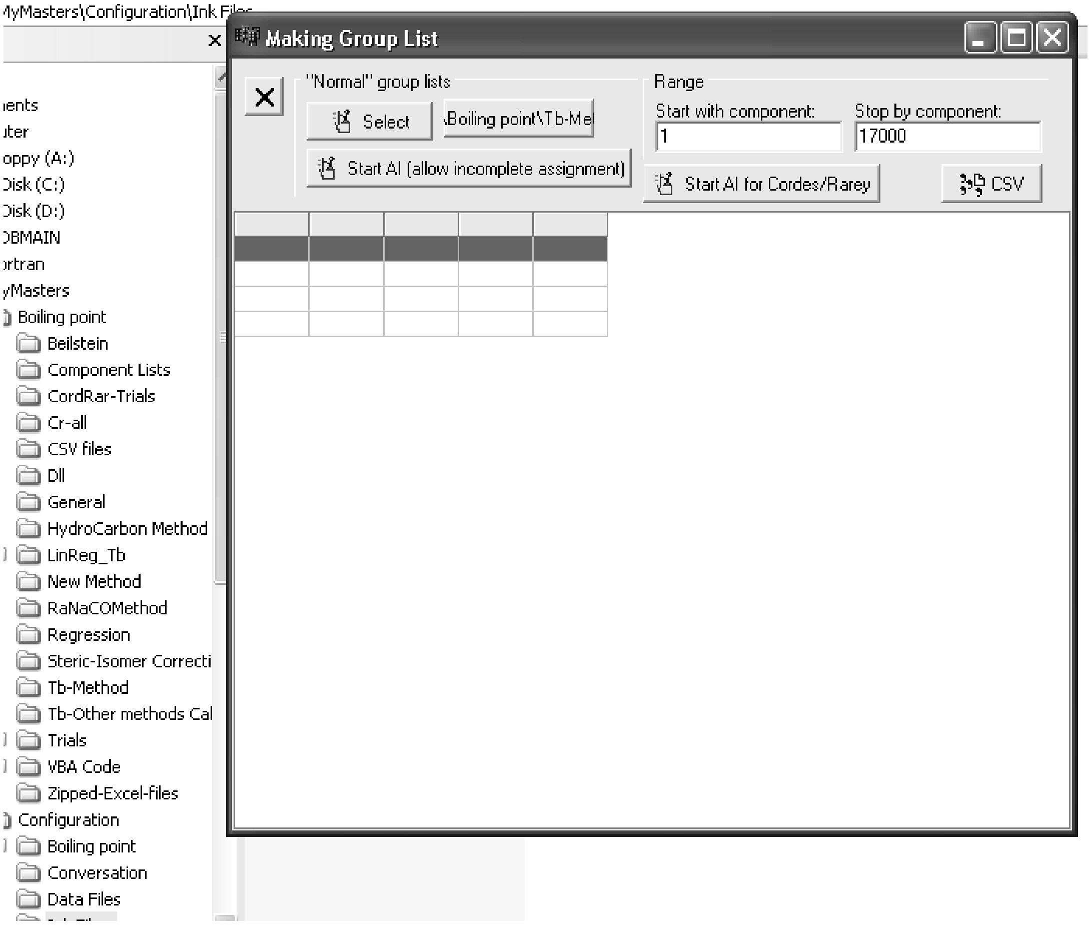
General notes:

- The number of atoms and number of bonds must correspond to the number of atoms and bonds description, for example, ‘4’ atoms correspond to line 3-6, ‘3’ bonds correspond to line 7-9.
- Hydrogen does not have to be included into the group; it is automatically calculated in the procedure. For example, the oxygen, number ‘3’ in Figure 4-7, has 1 maximum and minimum number of neighbours as defined in the ink file. However, it actually has 2 bonds, and thus the procedure will automatically assign hydrogen to the oxygen.
- * refers to all atoms, […] is used to group more than one atom, {…} is used to exclude atoms.
- The nomenclature of letters to describe the neighbourhood can be interpreted as follows: K – Chain, N – Non-aromatic, A – Aromatic, R – Ring and * - all neighbourhoods.

The only required input in the interface for the automatic fragmentation procedure is, as mentioned, the ink file, the start and end DDB components numbers. The procedure is an executable file, ‘MakingGroupList.exe’. The interface is presented in Figure 4-8. However, there are two important rules involving the fragmentation. Firstly, no atom can be assigned to more than one group, and, secondly, the entire molecule must be fragmented or the error 'group assignment failed' is recorded for the component.

Another automatic fragmentation procedure was developed, slightly modified from the procedure above, which allows for incomplete fragmentation of molecules. This is represented by 'Start Al (allow incomplete assignment)' in Figure 4-8. This procedure is useful for the fragmentation of second-order corrections. For this procedure, the two above rules do not apply.

The results of the both fragmentations are then saved as a comma separated variables (csv) file, which is easily imported into an Excel (xls) file.



Start with component: 1
Stop by component: 17000

Figure 4-8: Interface for the Automatic Fragmentation Procedure

4.3 Software Utilities and Terms

4.3.1 Development Platform

In the development of group contributions, the general requirements for analysis involve generating plots, tables and statistical analysis of groups of data. For this purpose, Microsoft Excel (MS-Excel) is used as the developmental platform for group contributions. It has many features, the most important being auto-filters. First of all, the structure of the worksheet is designed such that the columns contain the structural groups and the rows the components. The auto-filters can now be applied to each column. Thus, sets of filter settings can be stored and retrieved. The filter settings can be created and stored in a custom view. However, it has many limitations, especially when introducing, extending or changing different structural groups of the method. By hand, this can be time consuming and tedious work. However, Visual Basic for Applications (VBA), a powerful programming language, is easily integrated into MS-Excel, which can now be used as the user interface. The use of VBA has many advantages, which include:

- The ability to solve the regression for the Simplex algorithm and the least squares fit,
- program filter settings for a statistical analysis,
- perform simple calculations (on a set of 3000 data points and 200 parameters),
- ability to import data from text and structural files automatically,
- the use of object oriented programming,
- metalanguages, OLE servers, DLL files (all described below),
- etc.

4.3.2 Metalanguage

Meta is a prefix that in most information technology usages means "an underlying definition or description." Thus, *metalanguage* is a symbolic language used to describe and reason upon constructs of another programming language (base language). One could describe any computer programming or user interface as a metalanguage, for conversing with a computer.

The base language is generally a much more complicated language, for example, VBA. The metalanguage is a very simple programming language set in a user-friendly interface, for example, Microsoft Excel, whose commands are programmed in the base language. Thus it provides added flexibility and efficiency, as well as mobility, considering that the metalanguage has a more user-friendly interface than the base language. The metalanguage is of key importance to this work, since the time consuming problems associated with this project are essentially eliminated by the metalanguage.

4.3.3 Object Linking and Embedding (OLE)

OLE is primarily used to include objects from other components. These objects are typically documents or programs created by another component that supports OLE and are called OLE objects. A component that provides its documents or programs to be linked or embedded in other components is called an *OLE server*. A component in which documents or programs can be linked or embedded is called an *OLE container*. For example, a Microsoft Word document can be embedded in a Microsoft Access form and the user can then edit this document in Microsoft Word. In this case, Microsoft Word is the OLE server, and Microsoft Access is the OLE container.

The development or programming of an OLE is not of importance in this work; only a general understanding is needed, since it will provide access to objects embedded into certain components. In the DDB Artist program, there is a calculation toolbar which calculates properties of different methods. Thus, for a certain component, the boiling point can be calculated for different methods. This can provide as a comparison tool to the proposed method. Thus, the programmer of this tool was able to set OLE properties on the calculation component. This allows easy access from VBA, by defining the component as an object. The object only requires the DDB number of the component, the property and method names and the normal boiling point is then estimated. Thus, for all components and for all methods, boiling points can be easily estimated.

4.3.4 Dynamic Library Link (DLL)

A DLL is a file containing a collection of Windows functions designed to perform a specific class of operations. Most DLLs carry the .DLL extension, but some Windows DLLs, use the .EXE extension. Functions within DLLs are called by applications as necessary to perform the desired operation.

The importance of a DLL file in the current work can be more clearly seen by considering the problem. The non-linear regression, generally involving an average of 3 non-linear parameters, calls upon the least squares fit for optimized linear parameters. For every iteration, there are a number of linear optimisation calls, dependent on the Simplex algorithm (described earlier on). As a consequence, the time taken to perform a full regression, which takes about an average of 100-200 iterations, is approximately 18-36 hours. The actual time consuming operation of the full regression is the multi-linear least squares fit, since, the fit is performed on a set of about 3000 data points and 200 parameters. In VBA, this can take a considerable long time to fit, about 2 minutes. To solve this problem, a DLL file was developed in Compaq Visual Fortran, a programming language effectively much quicker in performing mathematical operations than VBA. The DLL incorporates the least squares fit, which is then called by VBA. Thus, in performing the least squares fit, the DLL is about ten times faster than VBA. This means that the full regression now only takes about 1-2 hours.

4.4 Construction of the Proposed Method Normal Boiling Point Estimation MS-Excel file.

The assumption for this construction is that the research (worksheets and program code development) have already been done. This construction is used for the purpose of the development and modification of structural groups. An additional advantage is that the construction can be easily adapted to any other property. For example, an MS-Excel file can be easily generated for the estimation of critical temperature and the research is made easier by the software development. Since there are a large number of worksheets and routine code, the procedure for the construction is developed to prevent any errors from occurring. The procedure also provides automation code routines, to allow a quicker and more efficient construction. The construction procedure is described in Appendix E. The procedure can also be referred to as an introductory manual to the reference files provided on the back cover of this book.

Chapter Five

Development of the Method

5.1 Introduction

In this work, one of the major aims was to derive a procedure or research strategy for the development of an estimation method for the prediction of pure component properties. Consider the use of the term 'estimation' in the above text. This term is not restricted to group contribution methods only, since there are a number of current estimation methods relating group contribution, molecular descriptors and molecular properties from molecular modelling. However, group contribution is the simplest form of estimation, but it has structural, physical and electronic limitations. These types of limitations or phenomena were described in Chapter 3. Consequently, the aim is to develop a group contribution method to its 'full capacity'. The term 'full capacity' can be defined as the maximum limit to which group contribution is able to perform estimations on selected components illustrating certain behaviour. Thus, the procedure chosen in this work involved performing a scientific statistical analysis of the different subclasses within a given chemical class of compounds. This analysis is achieved by the grouping of components or structural groups on a functional basis. Overall, the software developments (Chapter 4) are of key significance in the efficient, flexible and user-friendliness implementation of an estimation method. However, the first step is to develop a successful group contribution method for the prediction of the normal boiling point. This will provide a foundation for future estimation methods for all thermodynamic and physical properties.

From a review performed on various available group contribution methods (Chapter 2), the Cordes and Rarey (2002) method yielded the lowest average absolute deviation and probability of prediction failure. This is as a result of the method incorporating structural groups with a stronger relationship to the science of the normal boiling point. This method will serve as the basis for the proposed method. Thus, with the procedure described above incorporating an analysis of structural groups and components belonging to the specified group, the different phenomena, limitations and behaviour, if there are any, are detected, and resultant action, within the scope of group contribution, are taken.

5.2 Data Verification

The Dortmund Data Bank (DDB) includes normal boiling point temperature data for approximately 2800 components. The database has existed since 1973 and has been extensively used, for example, for the calculation of phase equilibrium data. Thus, the sets of data can be regarded as reliable. However, there are exceptions regarding the reliability of a pure component property. Consider an example for a set of 20 components. Assume that 19 of these components had no deviations and that 1 component has a deviation of 20K. The objective function would be the sum of squared deviation, which equals 400. This is similar to the above 20 components each having a deviation of 4.47K. Thus, errors in unreliable data are often greatly reduced by a simultaneous regression, which increases the deviation of the reliable data.

It must be presumed that there is a possibility of unreliable data in the database. Analyzing each component individually can be a tedious and time consuming procedure. Thus, the detection of unreliable data only involved components with extremely high deviations, for example, a component with an average absolute deviation greater than 15K. This led to the following errors occurring:

- Errors in the data file (nbp.dat). This file is used to import the experimental normal boiling temperatures into the Excel worksheet.
- Errors in the database. In some cases, there is a record of the normal boiling point in the data file, but not in the database.
- Unreliable sources of the experimental data. Some of these sources can also be outdated.
- Normal boiling points under-predicted or over-predicted from the extrapolation of low pressure data.
- Exotic components that cannot be captured entirely by group contribution estimations.

These components are then removed from the worksheet, and stored in 'R_data' worksheet.

5.3 Beilstein Database

The database contains normal boiling point temperatures for almost 19000 components. However, the methods of measuring the experimental data may not be of a precise or recommended type. For example, since the database is compiled by organic chemists, a component's boiling point may have been given as the temperature at which a component boils in a distillation column. Thus, the reliability of the experimental data is questionable. The data can serve as a test set, and more importantly, components can be added to groups where there are only few measurements available. These components can be also added to groups where there are no measured data available, but only with groups involving interaction parameters. In the case of components for which a new structural group would need to be defined, this would not be feasible because of its reliability.

The CAS registry is the largest and most current database of chemical substance information containing more than 22 million organic and inorganic substances and 36 million sequences. The CAS registry numbers have become the world standard and are not dependent upon any system of chemical nomenclature. They provide a reliable common link between the various nomenclature terms used to describe substances and serves as an international resource for chemical substance identifiers used by scientists, industry, and regulatory bodies.

In order to construct the Beilstein data set, the first objective is to obtain molecular structures for these components. The National Cancer Institute (NCI) has molecular structures for almost 250 000 components. However, there are only 122 672 components with CAS registry numbers. The CAS registry numbers represents a common link between the above databases. The procedure for obtaining a component's normal boiling temperature and molecular structure is described in Figure 5-1. The term 'Stoff' is a file which stores the component's name and other types of information, for example, molecular weight. The selected component with a normal boiling point from Beilstein and molecular structure from NCI is added to the private DDB, and is assigned a negative number.

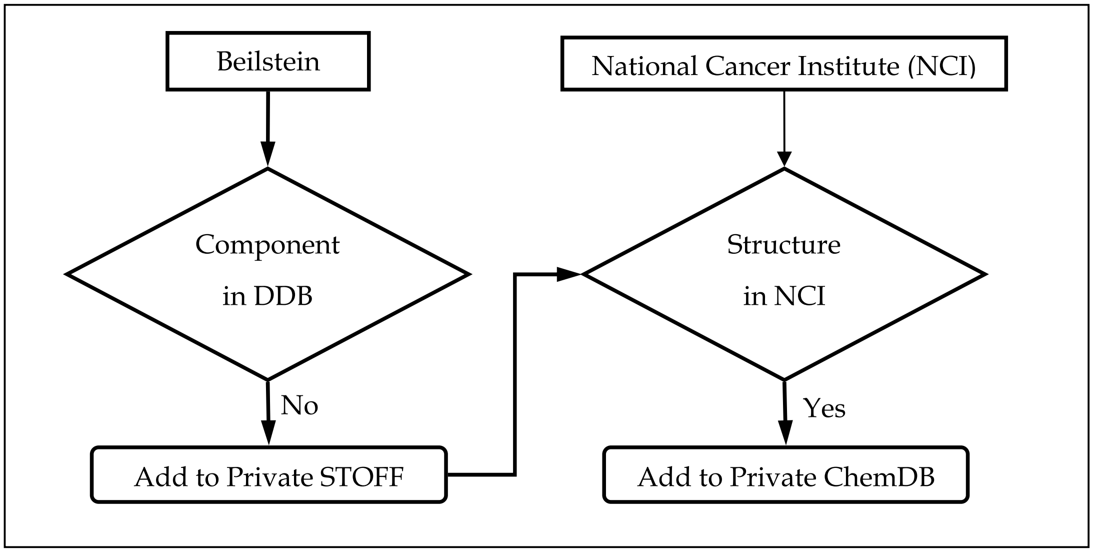


Figure 5-1: Simple Flow Diagram for the Retrieval of Normal Boiling points from the Beilstein Database

The extended data set can now be imported into the MS-Excel worksheet ('Tb-method cd-version.xls'). The fragmentation of components for which the molecular structures are available in the DDB (Public DDB), have already been carried out (Section 4.4). Thus, only the normal boiling points need to be imported into the worksheet. For the case of the private DDB, these components are fragmented (Section 4.2) and, together with the normal boiling points, are imported into the worksheet. In this manner, the extended data set now comprises of 1236 components from the Beilstein database, of which 1010 components were obtained from the private DDB. This set is based upon components for which structural groups are already defined.

5.4. Filter Program

The common problem associated with group contribution methods is the inability of the researcher to view the different types of phenomena or compound behaviour occurring, such as steric hindrance (Section 3.8). The developmental platform, MS-Excel and VBA (Section 4.3.1) provides auto-filters which can store functional groups into filter settings. A filter setting is a collection of certain column auto-filters which are specified in such a way, as to allow viewing of a specific functional group. For example, a hydrocarbon filter setting involves a collection of only hydrocarbon structural groups and is specified to allow viewing of only hydrocarbon components. These filter settings can be coded in VBA and programmed to generate results of specific functional groups. While this feature is important, there are many limitations, especially when dealing with over 100 filter settings and pages of code. The major limitation is that the structural groups are restricted to the code in VBA. This led to the development of a metalanguage (Section 4.3.2) filter program. The metalanguage is designed in MS-Excel, thus the filter criteria can be stored with almost unlimited complexity. VBA is used as the base language, which has an additional advantage of constructing and editing the filter program with simple code. The description of the filter program is presented in Appendix F.

5.5 Development of the Group Contribution Method - Part I: Mono-functional Compounds

An improved definition of mono-functional compounds is: a set of compounds with a hydrocarbon backbone and only one type of functional group, for example OH, NH_2, etc, which has a frequency of one. Thus the analysis is first performed on mono-functional compounds and different types of hydrocarbons, for example, n-alkanes. The approach is to analyse the performance of each group and test the predictive capability. The group definition, description, identification number (ID), priority (PR) and examples, for first-order groups and second-order corrections can be found in Tables B-1 and B-2 in Appendix B, respectively. For the proposed method, a detailed procedure is provided for the calculation of four different components in Tables C1, C2, C3 and C4 in Appendix C.

It should be noted that the research strategy suggested in this chapter is not performed only once, but a number of times. The first step was the analysis of the Cordes and Rarey method. The research strategy involved the scientific analysis of each functional group or subclass of compounds, as compared to previous methods. For each subclass of compounds, components with high deviations were detected and a solution was sought out. With the implementation of these improvements, the proposed method was then developed, which involved the construction of the new method (Section 4.4). The development was iterative, consisting of a scientific analysis, accompanied by the necessary modifications, implementation of the proposed method and so forth. Eventually, when there was no room for improvement within the scope of group contribution, the proposed method was finalised. Thus, the development or modification of groups or corrections introduced in this chapter is not in a time pattern of how the proposed method was developed. For example, most of the second-order corrections were introduced at the end of the analysis, but in this chapter, it will be introduced according to its chemical class. This also applies to structural groups and group interactions. The tables that follow in this section are provided to give the reader an idea of how the analysis was performed. Using the example of the introduction of a steric correction, which applies to hydrocarbons; this meant that since hydrocarbons are the backbone of all compounds, all other chemical classes were affected.

The regression of individual groups (Section 4.1.2) allows the researcher to test the performance of only that group. This allows the researcher to investigate the method in more detail to detect possible weaknesses. However in this work, the average absolute deviations of components presented involve a regression undertaken on all components. The available methods, with the exception of Stein and Brown, used extremely small data sets, usually involving a few or no multi-functional components of different chemical classes. This means that the regression favours the estimation of mono-functional compounds. Consequently, their estimations of multi-functional components have extremely high deviations. Thus in a few cases, the development of the proposed method involving mono-functional compounds have a slightly worse estimation than the available methods.

5.5.1 Hydrocarbons

Hydrocarbons represent the backbone of all organics compounds. Thus the development of the proposed method must involve hydrocarbons as the starting compounds. Table 5-1 presents deviations for the different types of hydrocarbons for the available group contribution methods.

Overall, the Cordes and Rarey method is far more accurate than other methods with a deviation of 6.89K for hydrocarbons. However, on a more detailed analysis, there were certain problems associated with alkene compounds. An error in the group definition of unsaturated non-aromatic hydrocarbons was detected. This error lead to the incorrect fragmentation of cumulated alkenes (C=C=C) as C=C-C. Thus a new group C=C=C (ID – 87) was introduced. In addition, conjugated alkenes, C=C-C=C (chain, ID – 89) C=C-C=C (ring, ID – 88), and conjugated alkynes C≡C-C≡C- (ID – 95) were introduced into the proposed method.

Table 5-1: Functional analysis of hydrocarbons showing the deviations and number of components for the different models used.

	Number of Components					Absolute Average Deviation (K)				
Compounds	JR	SB	GC	MP	CR	JR	SB	GC	MP	CR
Hydrocarbons	680	679	686	555	688	18.88	10.68	8.85	10.79	6.89
Saturated HC	266	266	266	266	266	20.05	14.18	7.87	9.22	6.63
n-Alkanes	27	27	27	27	27	55.69	12.10	18.67	13.64	6.47
Alkanes (non-cyclic)	192	192	192	192	192	25.14	16.36	8.49	7.18	6.68
Alkanes (cyclic)	74	74	74	74	74	6.85	8.53	6.26	14.52	6.51
Aromatic	177	167	177	115	177	29.12	7.27	12.04	9.03	6.70
Alkenes	173	180	180	126	180	9.18	8.53	6.16	17.54	7.42
Alkenes (cyclic)	49	53	53	26	53	6.92	6.42	6.16	15.13	8.13
Alkynes	35	35	35	33	35	13.13	12.16	13.84	3.98	5.40

(CR – Cordes and Rarey, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo)

The major disadvantage of the Cordes and Rarey method has been its inability to differentiate between isomers. On a detailed analysis of hydrocarbon compounds, there were rather high deviations for highly branched hydrocarbon isomers. For saturated hydrocarbons and non-cyclic saturated hydrocarbons, the previous method used only 7 and 4 group parameters respectively. The methods of Constantinou and Gani and Marrero and Pardillo used 16 and 17 group parameters for saturated hydrocarbons, and 9 and 10 group parameters for non-cyclic saturated hydrocarbons, respectively. The larger number of parameters in the latter two methods provided a better differentiation between isomer compounds. Thus, one way to improve the proposed method would be the introduction of several larger groups. But, the goal was not only to be as economical as possible, but scientific as well. The introduction of larger groups can greatly reduce deviations for certain components in the available database. However, this may lead to larger deviations for new components, which would then require the introduction of another new group. Thus, for these types of methods, the predictive capability of the method is now uncertain.

A common way to introduce information on hydrocarbon compounds is the use of topological indices. Ambrose (Reid et al. (1987)) introduced the Delta Platt number in the estimation of critical properties. However, the physical significance of these parameters is not only difficult to understand, but also even more complicated to generate. In general, these types of indices have relatively little or no relevance to the normal boiling point (consider the method of Wen and Qiang which uses a complicated GVS indice, but produces a similar average absolute deviation to the method of Constantinou and Gani).

The idea of the Delta Platt number is that it counts the number of carbons that are three bonds apart. Thus, to describe isomer effects, a group or correction must be able to provide information about the greater neighbourhood of the carbon atom. The groups defined in the proposed method contain information on one carbon atom only. However, the real differentiation for the large isomer deviations in the estimation methods deals with the hydrocarbon backbone. The general theory behind isomer differentiation has been discussed earlier (Section 3.8). Thus, from the above theory it became apparent that a correction needed to be introduced describing the effect of a C-C bond. The analysis of the regression results, led to the introduction of a steric hindrance and isomer correction. This correction involves the number of neighbours around a C-C bond. Thus, the molecule is sterically hindered if it contains more than 4 carbon neighbours. An example of a C-C bond with six neighbours is provided in Figure 5-2.

Since a carbon atom will generally have four neighbours, it is assumed that only a carbon functional group connected to it, will contribute to the steric hindrance of the molecule. Oxygen and nitrogen were, however, introduced as a steric parameter, but the regression produced a resultant improvement which was negligible. Also, in the case of a C-C bond with 3 neighbours on one carbon and 1 neighbour on the other carbon (C_3C-CC), no similar steric effect was found. The steric and isomer correction was also introduced for unsaturated hydrocarbons. In this case, only 1 neighbour may have a double bond, which produces a maximum of 5 neighbours. For the case of a 4 neighbour unsaturated correction (C=CC_3), the regression also produced no steric effect.

This presents an important point that if a collection of components illustrates similar physical or electronic phenomena, then a correction can be introduced to account for this behaviour. However, the difficult part is developing the correction within the bounds of group contribution. In the three cases above, it is seen that testing of these corrections led to no improvement. This can be attributed to there being no steric effect between the molecules.

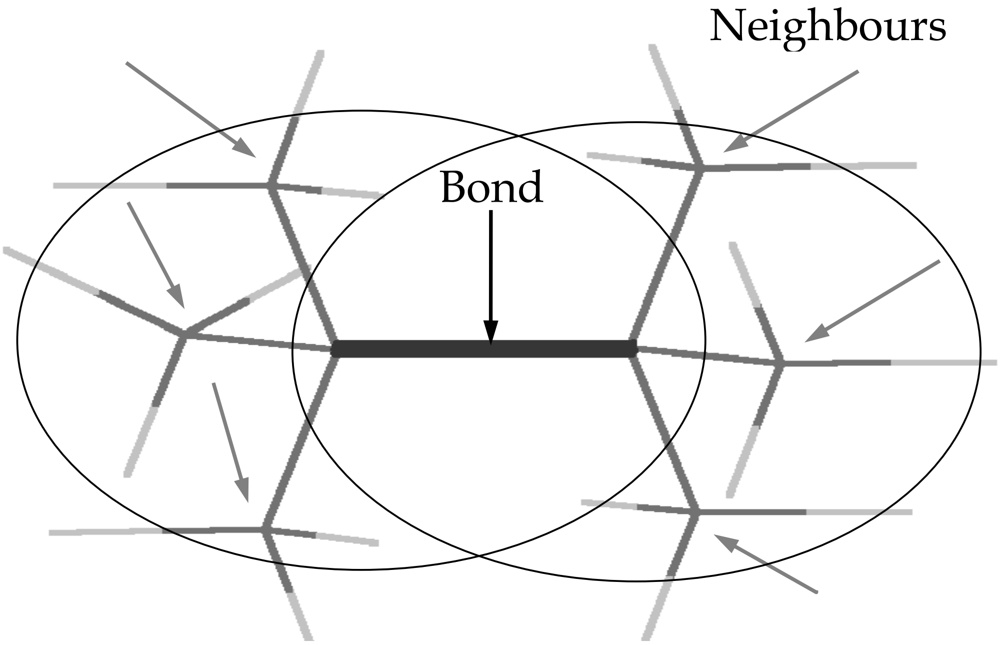


Figure 5-2: Steric and isomer contribution from the number of carbon atoms around a C-C bond.

For saturated hydrocarbons, 3 steric corrections were introduced, C_2C-CC_2 (ID – 131), C_2C-CC_3 (ID – 132) and C_3C-CC_3 (ID – 133). For unsaturated hydrocarbons, 1 steric correction was introduced, (C=)(C)C-CC_2 (ID – 130). The descriptions of these groups are presented in Table B-2 in Appendix B. An example for the estimation of 3,3,4,4-

tetramethylhexane, which includes a steric correction, is presented in Table C-1 in Appendix C.

5.5.2 Oxygenated Compounds

5.5.2.1 Alcohols

The estimation of the normal boiling point of alcohol compounds have always been a difficulty in previous methods. Stein and Brown included 4 and Cordes and Rarey 5 groups for the definition of these types of compounds. This enabled a much better prediction of mono-functional alcohols (Table 5-2). However, with multi-functional alcohols the prediction becomes even more complex. Consider the estimations of diol and triol compounds in Table 5-2. Subsequently, these high deviations will affect the regression involving mono-functional compounds. These types of compounds will be discussed later.

Table 5-2: Functional analysis of oxygenated compounds showing the deviations and number of components for the different models used.

	Number of Components					Absolute Average Deviation (K)				
Compounds	**JR**	**SB**	**GC**	**MP**	**CR**	**JR**	**SB**	**GC**	**MP**	**CR**
All alcohols	150	148	150	131	150	25.57	6.57	12.35	13.24	6.77
1-Alcohols	18	18	18	18	18	32.42	6.97	9.61	18.60	9.93
Primary alcohols	46	46	46	43	46	20.29	6.78	9.61	14.16	7.29
Secondary alcohols	45	44	45	44	45	25.54	5.32	11.54	13.19	5.85
Tertiary alcohols	31	31	31	18	31	39.97	6.76	19.88	17.90	6.65
Aromatic alcohols	28	27	28	26	28	18.35	8.03	9.83	8.55	7.52
Diols, Triols	22	22	22	22	22	32.05	26.78	28.12	18.79	16.79

(CR – Cordes and Rarey, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo)

The extra group for the Cordes and Rarey method incorporated a short chain alcohol group. This group was only applied to chains with less than five carbon atoms. However, this group included secondary alcohols and chains on aromatic rings. For short chain alcohol compounds, the effect of intermolecular forces, dipole moments and acid/base behaviour is more significant due to the smaller mass and size of the molecule. The distinction of deviations between the lower alcohol and higher alcohol components can be clearly seen in Figure 5-3. This implies that the short chain group on an aromatic ring should not be part of the group definition. Further testing resulted in the group being confined to components less than five carbons and limited to only primary alcohols. The description of the short chain group (ID – 36), primary (ID – 35), secondary (ID – 34), tertiary (ID – 33) and aromatic (ID – 37) alcohol groups are presented in Table B-1 in Appendix B.

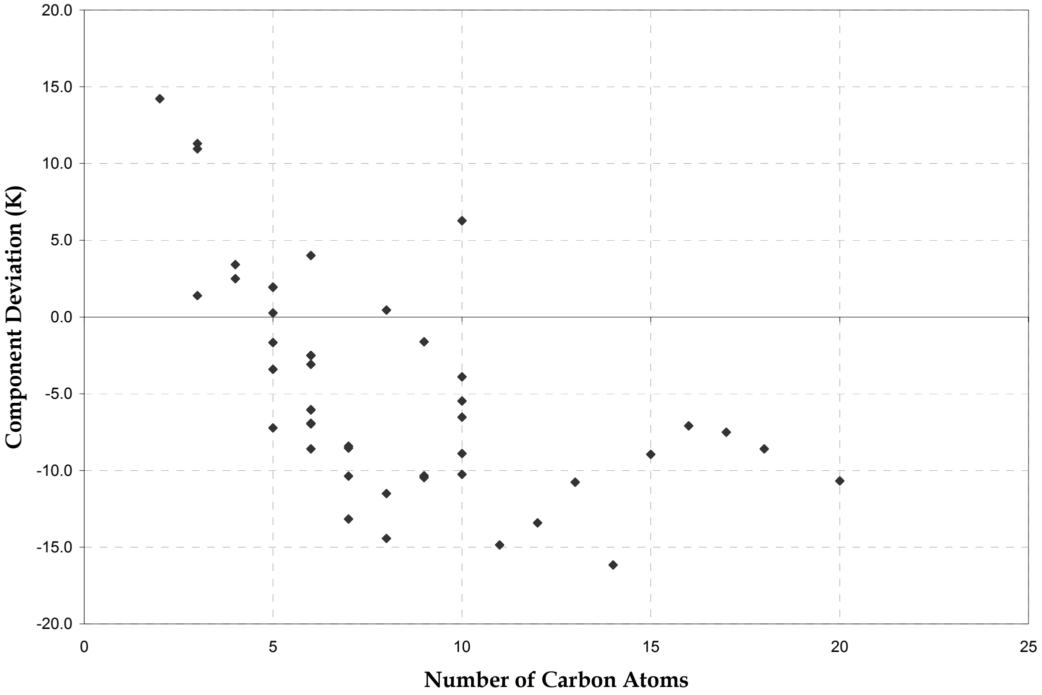


Figure 5-3: Component deviations (Cordes and Rarey) as a function of number of atoms for mono-functional alcohols.

5.5.2.2 Carbonyl Compounds

Carbonyl compounds (>C=O) are similar to alcohol compounds, in that both are highly electronegative and strongly influenced by intermolecular forces. There are four classes of carbonyl groups viz. aldehydes, ketones, esters and acids. In the Cordes and Rarey method, esters had been classified into an ester group (ID – 45), carbonyl di-ester (ID –

79), formic acid ester (ID – 46) and lactones (ID - 47). In conjunction with carboxylic acids (ID – 44) and acid chlorides (ID – 77), the analysis of these estimations of mono-functional compounds revealed no further classification. This can also be observed by the results produced in Table 5-3.

Table 5-3: Functional analysis of carbonyl compounds showing the deviations and number of components for the different models used.

	Number of Components					Absolute Average Deviation (K)				
Compounds	**JR**	**SB**	**GC**	**MP**	**CR**	**JR**	**SB**	**GC**	**MP**	**CR**
Aldehydes	26	26	26	20	26	13.93	7.79	9.52	7.69	8.41
Ketones	60	65	57	41	66	13.63	8.55	8.07	5.68	7.36
Carboxylic acids	34	34	34	31	34	34.70	7.05	13.44	12.33	6.69
Non-cyclic carbonates	4	4	4	4	4	25.07	49.74	3.46	5.50	3.65
Esters	110	110	103	87	110	16.60	4.27	7.68	6.83	5.27
Formic acid esters	0	17	17	13	17	-	64.48	8.71	2.43	4.82
Lactones	3	3	2	0	3	107.3	20.07	103.9	-	4.32
Acid Chloride	12	16	0	0	16	10.95	55.34	-	-	3.97

(CR – Cordes and Rarey, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo)

Aldehyde and ketone components for the previous method revealed higher deviations as compared to the available methods (SB and MP in Table 5-3). On an analysis of these components, the neighbouring carbon of aldehyde and ketone structural groups included both chain and aromatic structures. For the proposed method, two different aldehyde (ID – 52 & 90) and ketone (ID – 51 & 92) functional groups have been implemented, depending on whether the neighbouring carbon is part of an aromatic system or not.

As with alcohols, the description of multi-functional carbonyl components is more complicated and will be discussed later on. However, there are two carbonyl halogen corrections that have been included in the proposed method. These corrections will be described in multi-functional compounds, later on. There were large deviations observed for molecules with a carbon-carbon π-bond in conjugation with the carbonyl double bond. Thus, a common correction for the structure C=C-C=O (ID – 118) was introduced describing all four classes of compounds. It is not of importance, whether the C=C bond is part of an aromatic system or not. The correction takes into account the ability of the electronegative oxygen to polarize the electrons in the conjugated system resulting in a significantly larger charge separation than in the case of the isolated carbonyl double bond. An example for the estimation of methyl m-toluate, which includes a C=C-C=O correction, is presented in Table C-4 in Appendix C.

5.5.2.3 Other Oxygenated Compounds

These compounds involve ethers, epoxides and anhydride groups. The results of these groups are presented in Table 5-4. In the case of ethers (ID – 38 & 65) and epoxides (ID – 39), an analysis of these components produced no further improvements. The anhydride group, however, for a set of 7 components, produced an average absolute deviation of 25.4K. The cause of this high deviation is in the classification of the group in the Cordes and Rarey method, since all 7 components were represented by one group. This case is quite similar to the classification of aldehydes, etc. Since there are only 7 components, the component structures and deviations are represented in Figure 5-4. Thus, it can be clearly seen that the group proposed by Cordes and Rarey did not differentiate between chain and cyclic groups. Thus a new group was included describing cyclic anhydrides (ID – 96) containing an aromatic or double bond carbon in the ring (connected to a sp^2 carbon). The previous group was modified to comprise of only chain anhydrides (ID – 76).

Table 5-4: Functional analysis of ethers, epoxides and anhydrides showing the deviations and number for components of the different models used.

	Number of Components					Absolute Average Deviation (K)				
Compounds	**JR**	**SB**	**GC**	**MP**	**CR**	**JR**	**SB**	**GC**	**MP**	**CR**
Ethers	97	98	78	77	98	9.90	6.61	8.93	5.22	5.78
Aromatic oxygen	5	0	0	2	5	22.18	-	-	8.61	7.36
Epoxides	12	12	10	9	12	12.07	9.61	28.29	5.62	7.12
Anhydrides	7.0	7.0	4.0	0.0	7.0	26.9	45.6	25.1	-	25.4

(CR – Cordes and Rarey, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo)

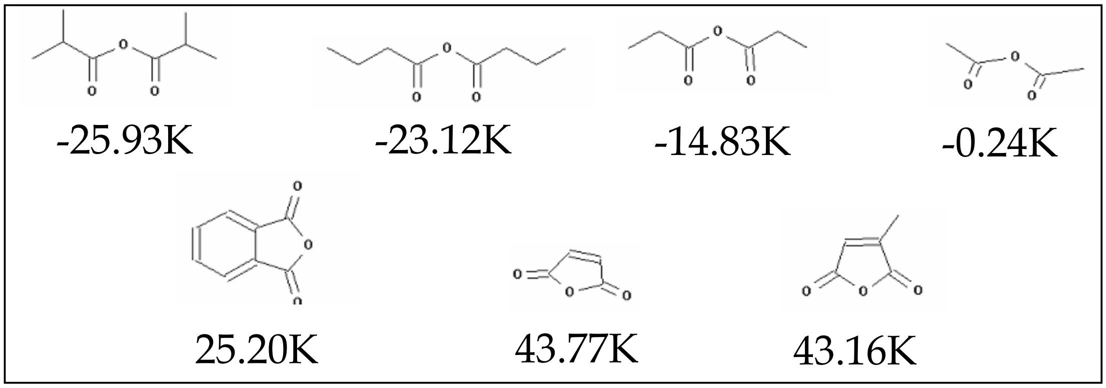


Figure 5-4: Component Structures and Deviations for Anhydrides.

5.5.3 Nitrogen Compounds

5.5.3.1 Amines

The Cordes and Rarey method included three different classes of amines compounds, namely primary, secondary and tertiary amines. The method also differentiated between non-aromatic (ID – 40) and aromatic (ID – 41) primary amines. Marrero and Pardillo produced the lowest deviation, however, with a rather small set of data and more parameters. The actual high deviations result from multi-functional amine components, which adversely affect the prediction of mono-functional components. From an analysis of amine compounds, the secondary amine group was differentiated between a non-aromatic (ID – 42), a ring (ID -97) and a neighbouring aromatic carbon (ID – 98) group. The tertiary amine was also differentiated between a non-aromatic group (ID – 43) and amine connected to a neighbouring aromatic carbon (ID -110) group.

Table 5-5: Functional analysis of amine compounds showing the deviations and number of components for the different models used.

	Number of Components					Absolute Average Deviation (K)				
Compounds	**JR**	**SB**	**GC**	**MP**	**CR**	**JR**	**SB**	**GC**	**MP**	**CR**
Primary amines	43	42	41	37	43	20.23	7.22	13.14	4.66	7.94
Secondary amines	36	36	35	30	36	14.06	10.35	12.37	5.65	8.86
Tertiary amines	18	18	18	16	18	17.46	10.68	9.08	8.00	7.95

(CR – Cordes and Rarey, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo)

5.5.3.2 Amides

An amide group consists of a carbonyl and an amine group. Table 5-6 presents the results of the amide compounds for available methods. With the exception of the Cordes and Rarey method, the available methods had extremely high deviations for these types of compounds. In some cases, these methods had a single or no group to describe this effect. Consequently, for the case of where there was no group defined, the fragmentation chose a combination of a carbonyl and amine group. The Cordes and Rarey method included 3 different groups for their estimation; amide (primary amine, ID – 50), mono-substituted amide (secondary amine, ID – 49) and disubstituted amide (tertiary amine, ID – 48). From an analysis of these types of components, there were no modifications made.

Table 5-6: Functional analysis of amide compounds showing the deviations and number of components for the different models used.

	Number of Components					Absolute Average Deviation (K)				
Compounds	**JR**	**SB**	**GC**	**MP**	**CR**	**JR**	**SB**	**GC**	**MP**	**CR**
Amides	3	4	3	0	4	111.0	47.77	8.04	-	8.58
Mono Amides	4	6	0	0	6	83.59	22.77	-	-	12.51
Di-Amides	2	8	5	0	8	74.45	11.38	19.16	-	6.80

(CR – Cordes and Rarey, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo)

5.5.3.3 Other Nitrogen Compounds

In this section there has been no development or modification involving these groups. The estimations of 5-membered N rings (ID – 66), 6-membered N rings (ID – 67), cyanides (ID – 57), isocyanates (ID – 80) and oximes (ID -75) are fairly accurate (Table 5-7). In the case of cyanides, the Stein and Brown method uses a non-ring and ring group. However, both of these groups have almost the same group contribution, implying that there need be no distinction between them. The high deviations seen in the Cordes and Rarey method, in most cases, are the result of multi-functional components.

The nitrite compounds considered in Table 5-7 are of two different types. The first is nitrous acid (-NO_2) and the second an ester of nitrous acid (ON-O-, ID – 74). The nitrous acid is classified into two groups, viz. neighbouring carbon attached to aliphatic (ID – 68) and aromatic carbons (ID – 69). Nitrates are the esters of nitric acid (ID – 72), which the available methods are not able to estimate.

Table 5-7: Functional analysis of nitrogen compounds showing the deviations and number of components for the different models used.

	Number of Components					Absolute Average Deviation (K)				
Compounds	**JR**	**SB**	**GC**	**MP**	**CR**	**JR**	**SB**	**GC**	**MP**	**CR**
5-membered N ring	9	12	0	1	12	44.52	29.27	-	0.11	7.14
6-membered N ring	28	28	20	14	28	13.20	7.52	21.60	4.13	8.99
Cyanides	26	26	19	23	26	22.42	3.38	7.91	8.45	10.09
Isocyanates	0	9	0	0	9	-	69.94	-	-	9.90
Oximes	0	0	0	0	9	-	-	-	-	7.88
Nitrites	8	8	7	1	15	28.81	12.37	6.03	0.01	7.82
Nitrates	0	0	0	0	5	-	-	-	-	3.83

(CR – Cordes and Rarey, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo)

5.5.4 Sulphur Compounds

In the analysis of sulphur compounds, there has been no development or modification of current groups. The deviations of the different types of functional groups, together with their ID number in parentheses, are presented in Table 5-8. The estimations of sulphur compounds are not as complicated as nitrogen and oxygen compounds. This is a result of sulphur not being a hydrogen bonding element. Hydrogen bonding is the strongest and most influential intermolecular force. Thus, the deviations presented in Table 5-8 are not as high as compared to other chemical classes, when compared to all methods.

5.5.5 Halogenated Compounds

Halogens consists of the following elements, F, Cl, Br and I. Halogens are not as electronegative as oxygen or nitrogen groups (for example OH, NH_2) but occur more frequently in a molecule. Thus, halogens contribute mostly to the dipole moment of the molecule. The mono-functional definition of halogenated compounds has been modified. Previously, the definition only included a frequency of one for a specific functional group. For halogenated compounds, the frequency can now have a value of greater than or equal to one.

Table 5-8: Functional analysis of sulphur compounds showing the deviations and number of components for the different models used.

	Number of Components					Absolute Average Deviation (K)				
Compounds	JR	SB	GC	MP	CR	JR	SB	GC	MP	CR
Disulfides (55)	4	0	4	3	4	7.96	-	20.17	1.12	3.38
Thiols (53)	37	37	18	18	37	12.91	7.21	7.60	7.46	5.13
Thioether (54)	30	30	28	19	30	15.45	8.05	11.49	3.88	4.56
Aromatic Thioether (56)	10	0	8	3	10	11.04	-	8.68	0.25	5.17
Sulfolane (82)	0	0	0	0	3	-	-	-	-	12.11
Isothiocyanates (81)	0	0	0	0	3	-	-	-	-	8.61

(CR – Cordes and Rarey, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo)

The estimations of the previous methods are presented in Table 5-9. In all cases the Cordes and Rarey method yields the lowest deviation, which is probably due to the better differentiation in structural groups and the large set of data used in the regression. From all halogens, fluorine is probably the most complicated group. Although, chlorine is more electronegative, fluorine takes part in hydrogen bonding. Thus, their estimations are generally higher than other halogens. From the analysis, there has been one modification to the Cordes and Rarey method. A fluorine group connected to a carbon already substituted with fluorine or chlorine and two other atoms, has been modified into two separate groups (ID – 21 & 102). Overall, the halogens groups are represented in Table B-1 with reference to ID – 21 to 32 & 102.

Since the dipole moment has the most influence on halogenated compounds, a number of corrections were tried and tested. But, since the dipole moment is defined by the vector addition of individual bond dipole moments (Section 3.4.2), it is virtually impossible to capture this effect within the scope of group contribution. In this case, a molecular mechanics calculation is probably the best option. However, there have been two corrections introduced to represent the two extremes of the dipole moment of a compound. The first correction involves a carbon attached to three halogens (ID – 121), which represents the highest dipole moment. The second is a secondary carbon attached to two halogens (ID – 122). The latter correction involves a situation where, even though there are two halogens, there is no dipole moment based upon the vector addition of each dipole bond cancelling out. An example for the estimation of perfluoro-2-propanone, which includes a halogen correction, is presented in Table C-3 in Appendix C.

Table 5-9: Functional analysis of halogenated compounds showing the deviations and number for components of the different models used.

	Number of Components					**Absolute Average Deviation (K)**				
Compounds	**JR**	**SB**	**GC**	**MP**	**CR**	**JR**	**SB**	**GC**	**MP**	**CR**
Fluorinated saturated	64	64	64	63	64	18.53	59.89	25.62	14.25	7.35
Fluorinated	86	86	86	83	86	18.01	48.96	26.46	12.83	8.02
Chlorinated saturated	64	64	60	0	64	26.01	14.72	8.26	-	7.43
Chlorinated	117	117	95	0	117	21.36	14.05	9.36	-	6.62
Brominated saturated	31	31	31	0	31	14.76	8.82	6.91	-	6.34
Brominated	49	49	49	0	49	12.89	8.88	7.48	-	7.36
Iodinated	18	18	18	15	18	13.63	6.91	5.77	12.06	5.20

(CR – Cordes and Rarey, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo)

5.5.6 Other Elemental Compounds

The Cordes and Rarey method proposes the broadest range of applicability of organic compounds. These compounds involve phosphates (ID – 73), arsine (ID – 84), germanium (ID – 85 & 86), stannium (ID – 83), boron (ID – 78) and silicon (ID – 70, 71 & 93). These average deviations are presented in Table 5-10. It can be clearly seen that, with the exception of Stein and Brown for stannium and silicon, no method has been able to estimate these types of compounds. The analysis of these components revealed only a modification to the silicon connected to at least an oxygen, fluorine or chlorine group. This group was differentiated into two classes. The first is silicon connected to at least one oxygen (ID – 71) and, the second is a silicon connected to at least one fluorine or chlorine (ID – 93) group.

Table 5-10: Functional analysis of other elemental compounds showing the deviations and number for components of the different models used.

	Number of Components					Absolute Average Deviation (K)				
Compounds	**JR**	**SB**	**GC**	**MP**	**CR**	**JR**	**SB**	**GC**	**MP**	**CR**
Phosphates	0	0	0	0	4	-	-	-	-	6.98
Arsine	0	0	0	0	6	-	-	-	-	3.07
Germanium	0	0	0	0	1	-	-	-	-	15.81
Germanium & Cl_3	0	0	0	0	3	-	-	-	-	6.03
Stannium	0	3	0	0	3	-	2.42	-	-	1.30
Borates	0	0	0	0	7	-	-	-	-	4.25
Silicon	0	23	0	0	27	-	22.18	-	-	4.34
Silicon to O, F or Cl	0	0	0	0	47	-	-	-	-	9.10

(CR – Cordes and Rarey, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo)

5.5.7 New Groups

The Cordes and Rarey method was developed in 1999. Since then, there have been a number of new experimental normal boiling points added to the database. The aim was to introduce new groups to broaden the range of applicability of the method. This involved compiling a set of components where the fragmentation had failed. The next step was to search for the error ‘group assignment failed’. This error occurs for components that could not be fragmented by the groups defined. This set was then filtered for components with experimental normal boiling temperatures. The new set of components was then analyzed for new structural groups. These groups are presented in Table 5-11.

Table 5-11: Description of new groups introduced into the proposed method.

Periodic Group	ID	Structure	Comments
16	103	-OCOO-	cyclic carbonate
	109	>S(C=O)-	carbonyl connected to sulphur
	94	-O-O-	Peroxide
	104	$>SO_4$	Sulphate
	105	$-SO_2N<$	sulfon amide
	107	>S=O	Sulfoxide
	99	-O-CO-N<	Carbamate
	100	>N-CO-N<	Urea
	116	>Se<	Selenium
15	91	-N=(C,Si)	double bonded non-aromatic amine
	101	$C_2>N<C_2$	quaternary amine
	111	N-C≡N	cyanide connected to nitrogen
	115	-ON=	Isoazole
	108	(S)-C≡N	Thiocyanate
	106	..=CNC=NC=..	Imadizole
	113	>P<	Phosphine
13	117	>Al<	Aluminium

5.6 Development of the Group Contribution Method - Part II: Multi-functional Compounds

5.6.1 Group Interactions

In theory, physical and thermodynamic properties depend on the physical interaction between molecules, of which one of the most critical factors is group interaction. As was described earlier for the case of alkane diols, all group contribution methods more or less fail when estimating multi-functional compounds, with considerably high deviations occurring. These types of multi-functional components generally occur when there is more than one strongly associating functional group. For these types of groups, the assumption of simple additivity is no longer observed.

The concept of non-additivity for multi-functional compounds can be considered by examining an example. A set of hydrocarbon alcohol compounds with their normal deviations for different methods are presented in Table 5-12. In almost all cases, the estimations by all methods are extremely poor. Considering the Cordes and Rarey method estimations, there seems to be a trend for these components. With the exception of a few components, for example, 1,2-hexanediol which has an extremely higher dipole moment, the normal deviations are generally positive, or under predicted. For this case, the normal deviation is calculated by subtracting the estimation temperature from the experimental temperature. The probable causes of these deviations result from the intermolecular interactions of the strongly associating alcohol groups (Figure 5-5). These intermolecular interactions are derived from any of the intermolecular forces, particularly hydrogen bonding, and non-bonded acid/base interaction. Thus, to counteract this effect, a group interaction parameter, in this case for an OH-OH interaction, was introduced.

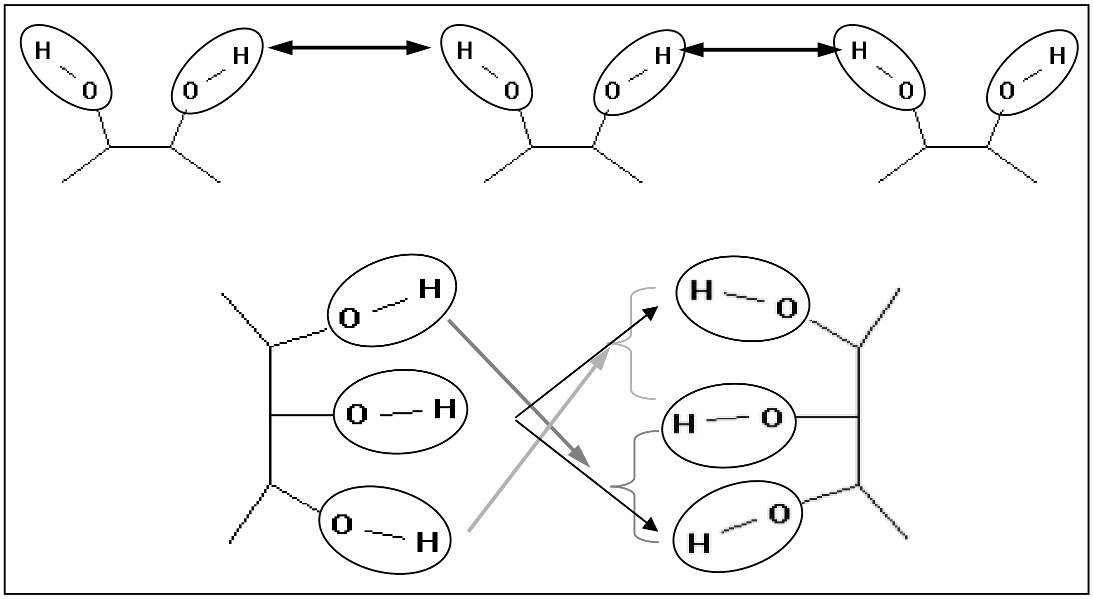


Figure 5-5: Group interaction for an alkane diol and triol.

Contrary to mono-functional compounds, the effect of group interaction decreases with the size of the molecule. This generalised effect is illustrated by the negative slope obtained in Figure 5-6. Thus, to take this effect into account, the sum of group interactions is divided by the number of atoms. There were also a number of tests performed to test this outcome; these tests are summarized in Table 5-13. The average absolute deviation is calculated using group interaction components only. These tests conclusively prove that the division by the total number of atoms (except hydrogen) attains the lowest deviation.

Table 5-12: Normal deviations and number of alcohol groups of models for multi-functional hydrocarbon alcohol compounds.

		Normal Deviations (K)				
Components	**A**	**JR**	**SB**	**GC**	**MP**	**CR**
1,2-Ethanediol	2	40.8	43.8	6.5	14.3	12.4
2,3-Butanediol	2	-18.5	26.2	16.1	-0.4	29.0
Glycerol	3	18.8	59.3	71.2	17.7	42.8
1,2-Propanediol	2	8.8	32.6	39.7	10.0	29.5
2-Methylpentan-2,4-diol	2	-46.9	16.6	5.5	-6.9	13.0
2-Butyne-1,4-diol	2	39.8	55.5	37.2	22.0	15.8
1,4-Butanediol	2	27.7	34.4	30.1	9.4	22.4
1,5-Pentanediol	2	17.3	27.3	24.8	3.6	16.0
1,3-Propanediol	2	35.4	39.8	34.5	12.9	27.3
1,6- Hexanediol	2	1.9	16.0	15.9	-6.8	5.7
3-Methyl-1,3-butanediol	2	-19.0	23.6	11.6	-2.0	15.3
1,3-Butanediol	2	6.7	32.0	27.6	10.5	27.8
2-Ethyl-2-Hydroxymethyl-1,3-propanediol	3	-42.0	14.2	21.5	-44.8	-3.5
2,2-Dimethyl-1,3-propanediol	2	-12.1	14.2	5.6	-19.8	2.7
2-Butene-1,4-diol	2	28.8	39.7	38.4	37.3	8.3
2,3-Dimethyl-2,3-butanediol	2	-67.3	11.6	-47.9	-34.0	2.8
3,4-Diethyl-3,4-Hexanediol	2	-99.5	-8.0	-55.4	-67.2	-14.0
1,2-Butanediol	2	-8.0	17.3	23.1	-4.1	13.1
1,2-Hexanediol	2	-51.2	-20.2	-13.3	-40.0	-24.5
Meso-erythrit	4	-55.0	36.3	65.6	-25.4	17.6
Cyclohexane-1,2-diol	2	-27.0	11.1	24.4	-2.9	23.1
2 Methyl-pentane-1,3- diol	2	-32.7	9.5	2.8	-21.4	2.8

(A – Number of alchol groups, CR – Cordes and Rarey, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo)

Table 5-13: Average absolute deviations for the different tests performed on group interaction components.

Tests	AAE
No Division	9.81
Division by total number of atoms (except hydrogen)	9.18
Division by number of carbon atoms	9.39
Division by number of carbon and nitrogen atoms	9.40
Division by total number of atoms except interaction groups	9.32

(AAE – Average Absolute Deviation (K))

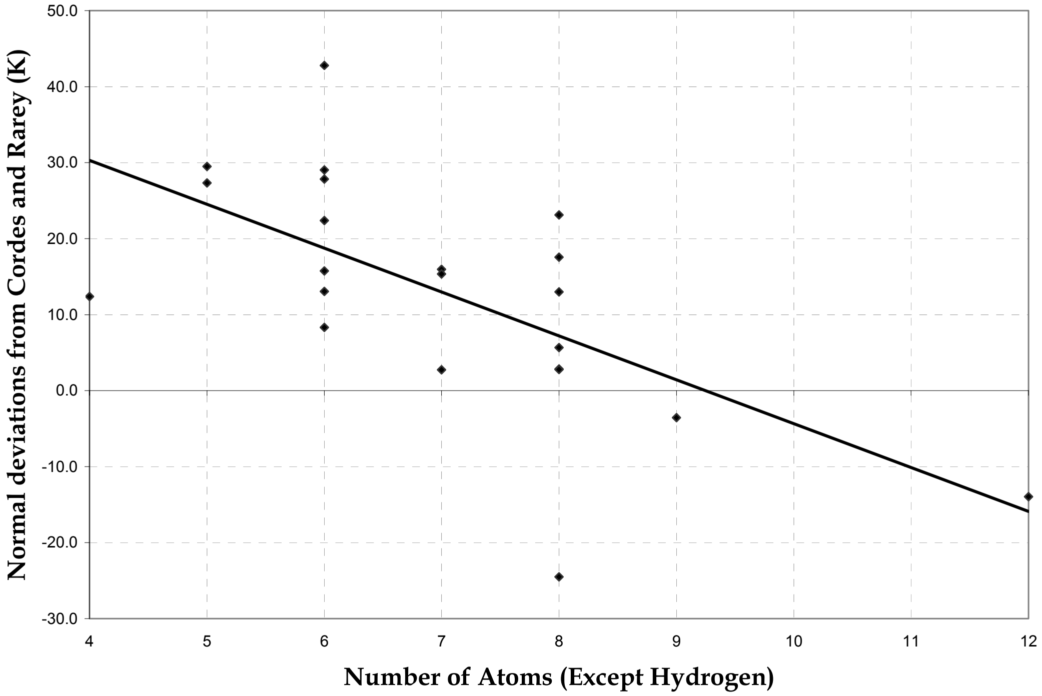


Figure 5-6: Graph illustrating the effect of number of atoms on normal deviations of hydrocarbon alcohols.

The interaction groups are derived depending on whether the group can act as an acid or base occurring in a component with a frequency greater than one. Thus, a search was performed to locate these interaction groups. In some cases, a combination of groups illustrating similar behaviour was used to denote an interaction group, for example, the non-aromatic alcohol interaction group is denoted by groups ID - 33, 34, 35 and 36. These interaction groups are presented in Table 5-14. Group interaction does not involve halogens, since these groups occur more frequently in components and cannot be captured by an interaction parameter.

Table 5-14: Groups considered to be non-additive (Structure and ID in parentheses)

A. Alcohol (-OH) (33,34,35,36)	B. Phenol (-OH(a)) (37)
C. Carboxylic Acid (-COOH) (44)	D. Ether (-O-) (38)
E. Epoxide (>(OC2)<) (39)	F. Ester (-COOC-) (45,46,47)
G. Ketone (-CO-) (51,92)	H. Aldehyde (-CHO) (52,90)
I. Aromatic Oxygen (-O(a)-) (65)	J. Thioether (-S(na)-) (54)
K. Aromatic Thioether (-S(a)-) (56)	L. Thiol (-SH) (53)
M. Primary Amine (-NH_2) (40, 41)	N. Secondary Amine (>NH) (42,97)
O. Isocyanate (-OCN) (80)	P. Cyanide (-CN) (57)
Q. Nitro (69)	R. Aromatic N in 5-ring (=N(a)-(r5)) (66)
S. Aromatic N in 6-ring (=N(a)-(r6)) (67)	

As opposed to simple additive groups, the frequency of a particular group interaction parameter A-B (for example, A,B = OH, NH_2 …) is calculated in a more complicated way. There are two important rules regarding the calculation of a group interaction parameter. Firstly, the interaction of the group with itself is already accounted for by the first-order group contribution parameter. Secondly, the group can only interact once with other interacting groups. The latter rule involves the possibility of interaction. For example, if an interaction group has a choice of interacting with two groups, then an equal possibility is chosen. This simply means that each interaction of a group with the other interacting groups has to be divided by the total number of interaction groups minus the interaction with itself. In case of two interacting groups A and B the total number of interactions is thus 2 (A-B and B-A, A-A and B-B were already accounted for by the first rule). As the parameters for A-B and B-A are identical, this gives $2 * C_{A-B} / (2 -1) = 2*C_{A-B}$. In case of 3 interacting groups A,B and C, there are in total 6 interactions (2 A-B, 2 A-C and 2 B-C) but each group can only interact with one of the two possible interaction groups at a time, so the sum of interaction contributions is $2*C_{A-B}/(3 -1) + 2*C_{A-C}/(3 – 1) + 2*C_{B-C}/(3 – 1)$. For example, a hydrocarbon triol (A-B-C) results in $3*C_{OH-OH}$; a glycerol monoester (2 OH groups, 1 ester group) results in $1*C_{OH-OH} + 2*C_{OH-ester}$. An example for the estimation of di-isopropanolamine, which includes a group interaction calculation, is presented in Table C-2 in Appendix C.

5.6.1.1 Group Interaction Metalanguage

The complexity involved in generating frequencies for group interaction parameters is best obtained by use of a metalanguage (Section 4.3.2). The group interaction metalanguage is quite similar to the filter language (Section 5.4) described previously, with MS-Excel as the interface and VBA as the base language. The description of the group interaction metalanguage is presented in Appendix G.

5.6.2 Carbonyl Halogen Correction

There are a large number of components in the data set with carbonyl groups comprising of halogens in close proximity. For these components, the dipole moment and polarisability is effectively increased. Thus, there are large deviations occurring for these types of components. This is the reason for two carbonyl halogen corrections introduced into the proposed method. The first corrections is for a carbonyl group connected to a carbon with two or more halogens (ID – 119) and second correction is for a carbonyl group connected to two carbons with two halogens each (ID – 120). An example for the estimation of di- perfluoro-2-propanone, which includes a carbonyl halogen correction, is presented in Table C-3 in Appendix C.

5.7 Development of the Group Contribution Method - Part III: Model Development

For the development of a relationship between the normal boiling point and group contribution, it is assumed that the relationship is purely mathematical. This relationship can be inferred from Figure 5-7. As a result, most of the earlier methods have tended to use a logarithmic fit for the prediction models. However, it can be clearly seen that there are a large number of outliers and the estimations for higher boiling point temperatures is less than expected. Consequently, this section will deal with the different mathematical approaches to fit this relationship.

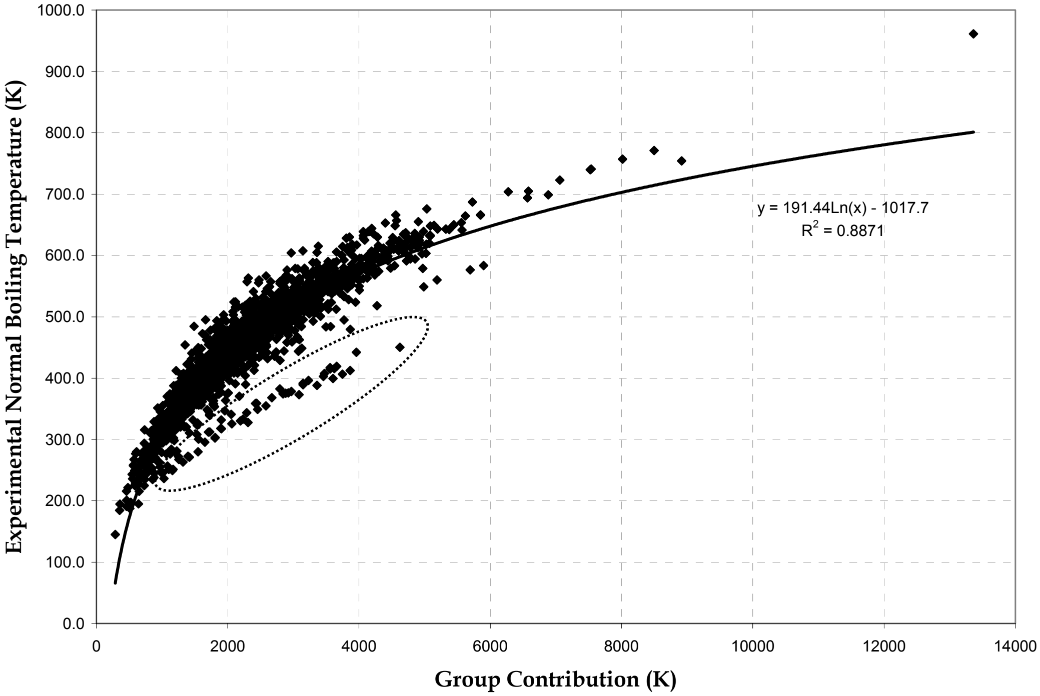


Figure 5-7: Normal boiling temperature as a function of the Cordes and Rarey group contribution value ($\Sigma N_i C_i$).

The model of Cordes and Rarey suggested that the relationship between the normal boiling point and group contribution is not a logarithmic fit (Equation 5-1), but rather dependant on the number of atoms (Equation 5-2). Since the molecular weight would provide additional information regarding the individual weights of each group, its inclusion, instead of the number of atoms, has to be physically tested (Equation 5-3). Also, the inclusion of both physical contributions was also tested, which can be integrated into two different models (Equation 5-4 & 5-5). The inclusion of molecular weight as a separate contribution was also tested (Equation 5-6). It is also apparent that the relationship in Figure 5-9 can be fitted to a power (Equation 5-7) and logarithmic (Equation 5-8) model. These models were also tested.

$$T_b = a\,(\ln \textstyle\sum_i N_i\, C_i) - b \qquad (5\text{-}1)$$

$$T_b = \frac{\sum N_i C_i}{n^a + b} + c \tag{5-2}$$

$$T_b = \frac{\sum N_i C_i}{M^a + b} + c \tag{5-3}$$

$$T_b = \frac{\sum N_i C_i + d}{en^a M^f + b} + c \tag{5-4}$$

$$T_b = \frac{\sum N_i C_i}{dn^a + eM^f + b} + c \tag{5-5}$$

$$T_b = \frac{\sum N_i C_i}{n^a + b} + cM^d + e \tag{5-6}$$

$$T_b = \frac{(\sum N_i C_i)^d - e}{n^a + b} + c \tag{5-7}$$

$$T_b = \frac{d(\ln \sum N_i C_i) + e}{n^a + b} + c \tag{5-8}$$

The model proposed by Marrero and Pardillo included molecular weight as a linear relation to the normal boiling point. This model has been modified to include, firstly the number of atoms (Equation 5-9), and then both contributions (Equation 5-10). Retzekas et al (2002) proposed a model including the physical parameters as a separate contribution for the estimation of only petroleum and coal liquid fraction hydrocarbons. The model also included the density as part of the physical contribution. In this case, the availability of experimental density data is limited, since a large number of components are supercritical. Thus, the model has been modified to exclude the density, with the inclusion of the number of atoms (Equation 5-11).

$$T_b = (\sum_i N_i C_i)^a n^b + c \tag{5-9}$$

$$T_b = (\sum_i N_i C_i)^a n^b M^c + d \tag{5-10}$$

$$T_b = a\left(\sum_i N_i C_i\right)^b + c\, M^d n^e + f \qquad (5\text{-}11)$$

The analysis of the above models will provide a means of assessment to test the importance of the physical contributions. Firstly, the relationship of these physical contributions to the normal boiling point and group contribution needs to be tested. Although the normal boiling point and group contribution are interrelated, their relationship to the physical contributions is different. Thus, if a better result is obtained with the physical contributions in the numerator, it can be assumed that the relationship is directly proportional to the normal boiling point. For example, an increase in the number of atoms produces an increase in the normal boiling point, consider Equation 3-11. On the other hand, if a better result is obtained with the physical contributions in the denominator, then the reverse applies, i.e. it is now dependent on the group contribution. The effect of including the physical contribution in both the numerator and denominator (Equation 5-12) was also tested. Secondly, the effect of including the number of atoms, molecular weight or both will also be tested.

$$T_b = \frac{M^d \sum N_i C_i + e}{n^a + b} + c \qquad (5\text{-}12)$$

Equations 5-13 to 5-16 were also tested involving different variations of the above models. Equations 5-17 and 5-18 represent a quadratic function on the normal boiling point and group contribution, respectively. The inclusion of a polynomial equation was not tested, since it is expected that this type of function would not be suitable for the above relationship.

$$T_b = \frac{\sum N_i C_i}{M^a + b} + \frac{\sum N_i C_i}{M^d} + c \qquad (5\text{-}13)$$

$$T_b = \frac{\sum N_i C_i}{M^a + b} + \frac{\sum N_i C_i}{M^e} + \frac{d}{M^c} + f \qquad (5\text{-}14)$$

$$T_b = \frac{\sum N_i C_i}{n^a + b} + \frac{\sum N_i C_i}{n^d} + c \qquad (5\text{-}15)$$

$$T_b = \frac{\sum N_i C_i}{n^a + b} + \frac{\sum N_i C_i}{n^e} + \frac{d}{n^c} + f \qquad (5\text{-}16)$$

$$dT_b^2 + eT_b + f = \frac{\sum N_i C_i}{n^a + b} + c \qquad (5\text{-}17)$$

$$T_b = a(\sum N_i C_i)^2 + b\sum N_i C_i + \frac{c}{n^d + e} + f \qquad (5\text{-}18)$$

The above models represent the various significant mathematical formulae used to describe the relationship between the normal boiling point and the total group contributions. However, it is obvious that different functional groups will behave differently according to the mathematical model being applied. For example, a regression performed on alcohol compounds will yield different non-linear parameter values than that of a regression on halogen compounds. Consider for example, the trend encircled in Figure 5-7. On an analysis of these components revealed a large number of halogen compounds. Thus, it is apparent that a regression performed on all compounds will yield optimised values to fit all functional groups. To try and account for this effect, a second set of contributions was developed. The regression of these groups is, however, much more complicated. Thus, a successive approximation was developed for this modified regression. This type of regression entails repeatedly optimising one or a few variables while keeping the other variables fixed. The algorithm for the successive approximation is presented in Figure 5-8. Implementing the second set of contributions, the first approach was based upon the number of atoms of each functional group (Equation 5-19). Following this approach, it was decided to also base the second set on the exponent of the number of atoms (Equation 5-20) and summation of group contributions (Equation 5-21). In the latter two cases, these models will provide different non-linear parameters for each functional group.

$$T_b = \frac{\sum N_i C_i}{(\sum B_i D_i)^a + b} + c \qquad (5\text{-}19)$$

$$T_b = \frac{\sum N_i C_i}{n^{a(1+\sum B_i D_i)} + b} + c \qquad (5\text{-}20)$$

$$T_b = \frac{\sum N_i C_i^{(1+\sum B_i D_i)}}{n^a + b} + c \tag{5-21}$$

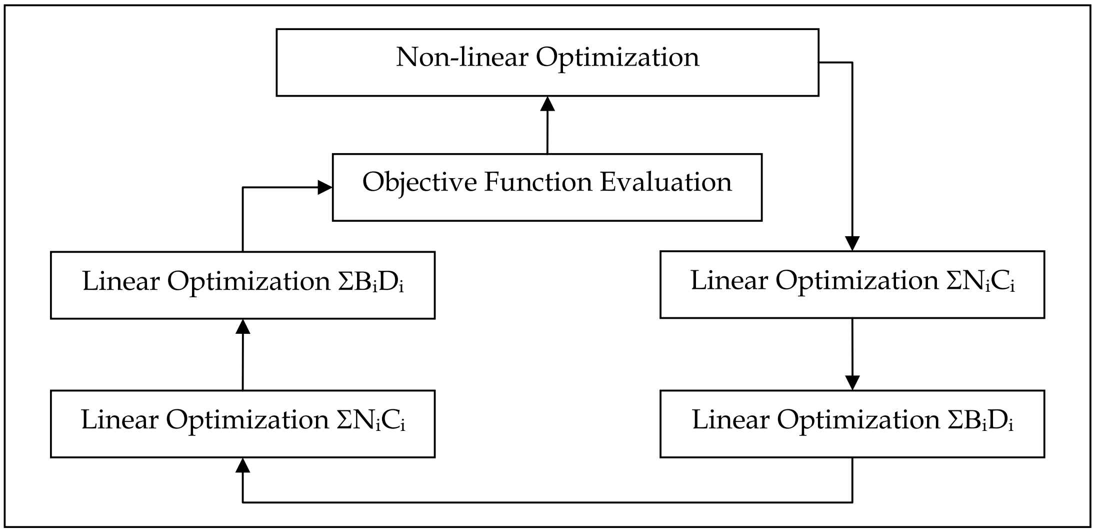

Non-linear Optimization
Objective Function Evaluation
Linear Optimization ΣBᵢDᵢ
Linear Optimization ΣNᵢCᵢ
Linear Optimization ΣNᵢCᵢ
Linear Optimization ΣBᵢDᵢ

Figure 5-8 Successive approximation algorithm, for the regression of a second set of contributions.

Chapter Six

Results and Discussion

This chapter analyzes the relevance of the results and examines the discrepancy between estimations and experimental data. The analysis will allow a detection of any possible weaknesses of the proposed method as compared to other methods and identify typical components for which extraordinarily large deviations occur. The results of the regression of the first-order groups, second-order corrections and group interactions are presented in Tables D-1, D-2 and D-3 in Appendix D. These groups are presented with respect to their ID numbers, together with the group deviations and number of components. The definition of each group is given in Appendix B.

6.1 Hydrocarbon Compounds

6.1.1 Mono-functional Hydrocarbons

It was discussed in Chapter 2 that previous group contribution methods have a propensity to provide more information about the molecule. But, this information practically provided very little or no significance to the boiling point. This view can be evidently observed in Table 6-1. In a few cases, their predictions are adequate, but in most cases the deviations are reasonably high. These results will be plainly seen in all types of compounds. Also, these poor predictions will lead to higher deviations involving other functional compounds. The whole idea of the research strategy is to develop a solution (in the form of a group or correction) to fit the different phenomena and behaviour (Chapter 3) that occurs in the many different types of functional compounds. This solution, however, must be within the scope of group contribution, in other words, can only be derived from the molecular structure of a compound. Consequently, the inclusion of molecular properties or molecular descriptors will be able to account for certain component behaviour out of the scope of group contribution.

Table 6-1: Functional analysis of hydrocarbon compounds showing the deviations and number of components for the different models used.

	Number of Components						Absolute Average Deviation (K)					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
Hydrocarbons	689	680	679	686	555	688	**5.80**	18.88	10.68	8.85	10.79	6.89
Saturated HC	266	266	266	266	266	266	**4.92**	20.05	14.18	7.87	9.22	6.63
Alkanes (non-cyclic)	192	192	192	192	192	192	**4.95**	25.14	16.36	8.49	7.18	6.68
Alkanes (cyclic)	74	74	74	74	74	74	**4.85**	6.85	8.53	6.26	14.52	6.51
Aromatic	177	177	167	177	115	177	**6.04**	29.12	7.27	12.04	9.03	6.70
Alkenes	180	173	180	180	126	180	**6.55**	9.18	8.53	6.16	17.54	7.42
Alkenes (cyclic)	53	49	53	53	26	53	**7.78**	6.92	6.42	6.16	15.13	8.13
Alkynes	35	35	35	35	33	35	**3.01**	13.13	12.16	13.84	3.98	5.40

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

An overall analysis of hydrocarbons revealed large deviations for components consisting solely of a single group such as benzene (17.1K), cyclopropane (23.9K), cyclobutane (18.4K), cyclopentane (12.7K) and cyclohexane (19.2K). These cyclic structures have a greater stability than larger cyclic structures. The above components also represent the first elements in a homologous series. There were also large deviations for ethane (8.4K) and propane (13.7K). It was discussed in Chapter 3 that these types of components, in most cases, disobey the general trend. Consequently, it is expected that these components would have higher deviations. Previous methods generally incorporated much smaller data sets, and the regression was performed on only a few components of a homologous series. In view of the fact that these compounds are readily available in common textbooks or databases, no special group or correction was introduced.

With the exception of cyclic alkenes, the proposed method has the lowest average absolute deviation for the different types of hydrocarbons. For cyclic alkenes, there were large deviations for the first few components of this series, like cyclopentene (19.5K) and cyclohexene (18.5K). There were also large deviations for components with two double bonds like 1,4-cyclohexadiene (15.3K). The effect of alkenes with respect to the molecular size and thermodynamic properties were discussed in Chapter 3.

Overall, the effect of a double bond is quite similar to an alcohol group, but the polarizability is much weaker. With a double bond, the sp^2 carbon has a stronger electron potential and the electron flow produces a dipole moment. Subsequently, when there are two double bonds, the effect is more complicated particularly with cyclic structures (a cyclic structure with a double bond now has a sp carbon). Thus, the prediction of boiling points for these compounds is quite complicated for a group contribution estimation method. Previous methods introduced larger groups; however, this will not improve the overall estimation, since it depends on the positioning of the double bonds in a molecule. The introduction of the larger group's scenario was discussed in Chapter 5. It should be emphasized that certain structural and electronic limitations of group contribution should not be corrected by the inclusion of larger groups. This will hinder the introduction of molecular properties.

The comparison for the different models for estimations for n-alkanes is presented in Table 6-2. The comparison also includes the method of Marrero and Gani (2001). Due to the complexity of the Marrero and Gani method, the calculation was done manually for n-alkanes and is given in the Microsoft Excel file 'n-alkanes for different models.xls' in the reference CD on the back cover of this thesis. The relationship between the estimated normal boiling points and molecular weight for the different methods is also presented in Figure 6-1.

Table 6-2: Functional analysis of n-alkanes showing the deviations and number of components of the different models used.

		Absolute Average Deviation (K)						
Compounds	**NC**[a]	**Pr**[b]	**JR**[c]	**SB**	**GC**	**MP**	**CR**	**MG**
n-Alkanes	27	**5.54**	55.69	12.10	18.67	13.64	6.47	19.99

[a] NC – Number of Components, [b] Proposed method, [c] JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey, MG – Marrero and Gani

The Marrero and Gani method produces an extremely high average deviation for the estimation of n-alkanes. The calculation for these compounds included no second or third-order groups, since there were none defined. Thus, it can be clearly seen that when the second or third-order groups are not applied, the predictions are poor. This emphasizes the point discussed in Chapter 5 that second and third-order groups were only included to reduce deviations of components already in the database.

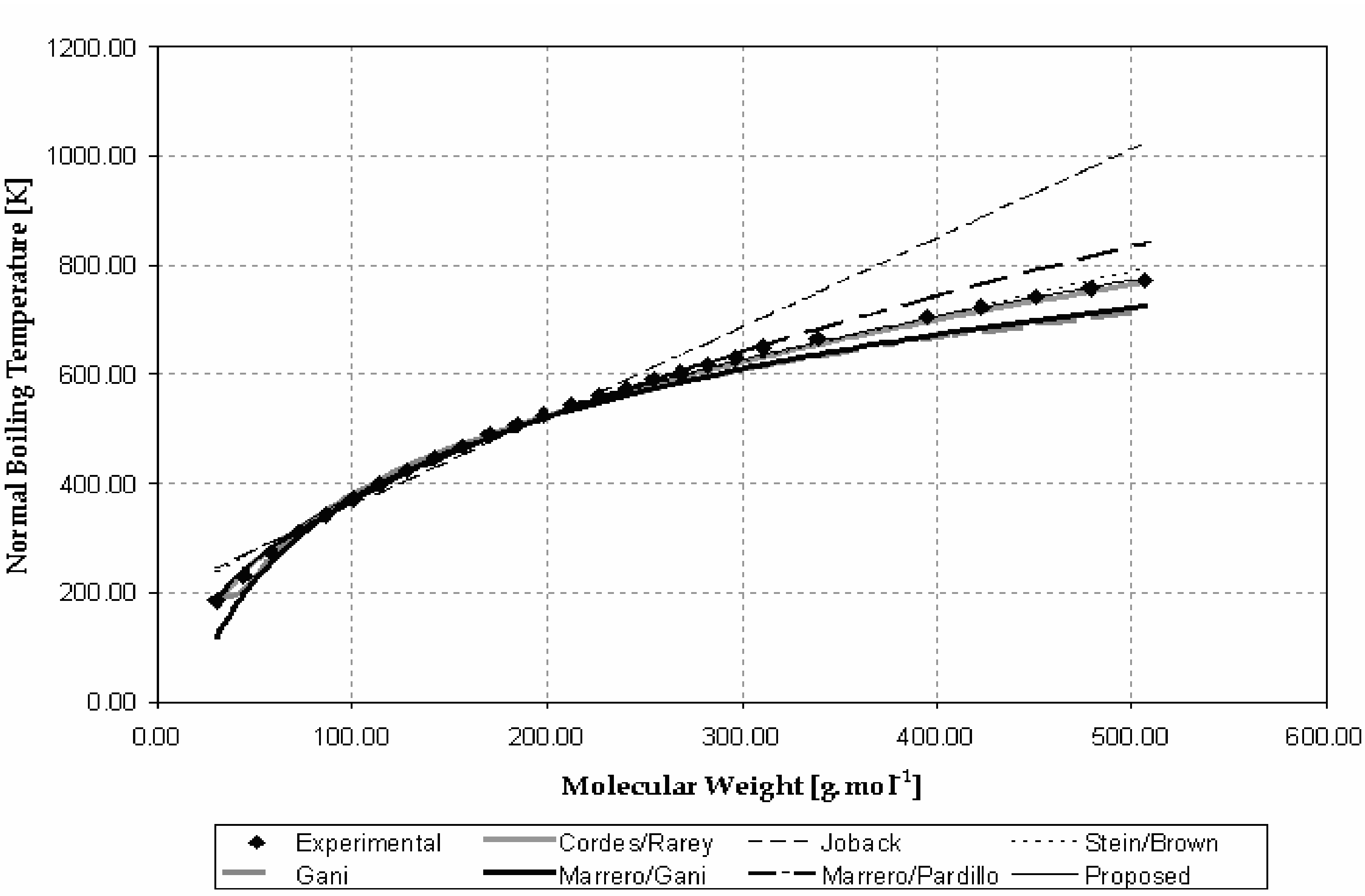


Figure 6-1: Normal boiling temperature vs. Molecular weight of n-alkanes for different models.

Consequently, the estimation of new components or predictive capability is uncertain. This can also be seen by the curve for the Marrero and Gani method in Figure 6-1. For higher molecular weight components, their predictions are extremely poor. Firstly, it can be considered that these components were probably not in database at the time of development (Marrero and Gani employs approximately 1100 components). Secondly, the curve flattens out as a result of the logarithmic fit used. For the Constantinou and Gani method, the curve also flattens out as a result of the logarithmic fit. Consequently, both methods are restricted when dealing with higher molecular weight components. The proposed method, however, yielded an excellent result in accordance with the experimental boiling temperatures. In Figure 6-1, the curve for the proposed method is slightly overlapped by the previous method of Cordes and Rarey.

Experimental data for heavier hydrocarbons is rather limited. These compounds can arise from, for example, byproducts from the processing of crude oil in the petroleum industry. Separation of these compounds involves measuring thermophysical properties which can be relatively expensive. In some cases, group contribution methods are used. Thus, the predictive capability of a group contribution method for the estimation of hydrocarbon compounds should be tested for the heavier components. Table 6-3 presents data for different types of hydrocarbons for carbon numbers greater than 19. The proposed method is far more accurate in all cases, with the exception of aromatic hydrocarbons as compared to the Cordes and Rarey method. For the case of unsaturated non-aromatic hydrocarbons, there are no experimental data in this range. These types of compounds are generally unstable and consequently decompose before the boiling point is reached.

6.1.2 Steric and Isomer Correction

The steric correction accounts for steric hindrance between C-C bonds. Table 6-4 presents the results for the four steric corrections introduced. In all cases, the proposed method produced far better results. This confirms the discussion in the previous chapter of collecting components illustrating similar physical phenomena. For example, C_3C-CC_3 (ID -133) is definitely the most significant of the corrections introduced. Consequently, the correction has a significantly higher group contribution value, which is obtained from the regression of all compounds (Appendix D). This point is also evident by the larger resultant improvement in the results. The introduction of the correction also indicated no significantly worse predictions for other components. Consequently, it can be concluded by introducing a scientific correction rather than a larger group, not only is the average absolute deviation lower, but also the predictive capability is improved.

Table 6-3: Functional analysis of hydrocarbon compounds for carbons greater than 19, showing the deviations and number of components for the different models used.

	Number of Components						Absolute Average Deviation (K)					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
Hydrocarbons	49	49	49	49	49	49	**6.47**	85.68	9.00	20.49	20.71	8.01
n-Alkanes	9	9	9	9	9	9	**5.05**	132.4	10.61	39.67	31.53	9.03
Alkanes (non-cyclic)	31	31	31	31	31	31	**6.65**	87.41	9.74	22.57	21.47	9.20
Aromatic	17	17	17	17	17	17	**6.23**	85.25	7.56	16.60	20.28	5.83

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

For the case of 2,3,3,4-tetramethylpentane (Figure 6-2), there are two $C_3C\text{-}CC_2$ steric corrections. However, both corrections have a common carbon (centre carbon in Figure 6-2). Because there were no other components depicting this type of scenario, there were no special corrections introduced for this component (-26.4K).

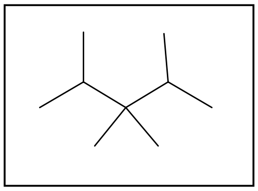


Figure 6-2: Molecular structure of 2,3,3,4-tetramethylpentane illustrating two steric corrections with a common carbon.

Overall, the introduction of the steric correction led to a better estimation for heavier hydrocarbons (Table 6-3). The correction was also introduced for the differentiation between isomers. Consider the estimation of C_7H_{16} to $C_{12}H_{26}$ and their isomers presented in Table 6-5. The proposed method yields a much better estimation than the previous method. The deviations are also in close range with the methods of Constantinou and Gani and Marrero and Pardillo, of which both methods use a larger number of groups (Section 5.5.1). It is also evident for the latter method, that as the compounds become heavier, the estimations worsen.

Table 6-4: Functional analysis of compounds involving steric corrections, showing the deviations and number of components for the different models used.

	Number of Components						**Absolute Average Deviation (K)**					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
$((C=)(C)C-CC_3)$	27	27	27	27	19	27	**6.59**	33.39	11.51	13.08	12.09	8.86
C_2C-CC_2	88	88	88	85	77	88	**5.37**	13.90	23.02	10.99	11.81	7.51
C_3C-CC_2	44	41	44	41	22	44	**4.41**	23.47	21.48	8.60	11.67	9.02
C_3C-CC_3	17	16	17	17	7	17	**3.23**	21.91	19.77	7.00	17.51	10.27

(Proposed method, [b] JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

Table 6-5: Boiling point deviations of C_7H_{16} to $C_{12}H_{26}$ compounds and their isomers.

		Absolute Average Deviation (K)					
Compounds	**NI**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
C_7H_{16}	9	**2.32**	5.17	13.05	2.79	2.71	3.62
C_8H_{18}	16	**2.71**	6.77	21.15	4.33	2.82	5.63
C_9H_{20}	29	**3.76**	7.65	24.64	4.75	2.84	6.75
$C_{10}H_{22}$	17	**3.57**	7.99	24.70	3.09	4.04	6.22
$C_{11}H_{24}$	9	**2.33**	9.83	17.49	3.44	4.93	2.52
$C_{12}H_{26}$	10	**3.54**	9.34	19.83	2.63	5.11	4.62

(Number of Isomers, Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

6.2 Oxygen Compounds

6.2.1 Alcohol Compounds

For the estimation of mono-functional alcohols, the proposed method yielded excellent results (Table 6-6). For the case of 1-alcohols, ethanol produced the highest deviation (18.4K) as compared to Stein and Brown (13.2K) and Cordes and Rarey (14.2). Since ethanol is the first component of the alcohol series, the result is as expected. Also for secondary and tertiary alcohols, the first components of these series also produced high deviations, 2-propanol (28.9K) and tert-butanol (20.4K), respectively. The complexity of alcohol components, especially with a rather small molecular weight can be compared to the normal boiling points of water and methane (Section 3.4.3). There were also large deviations observed for components with competing dipole bonds and polarisabilities (Figure 6-3), like 2-propyn-1-ol (21.5K).

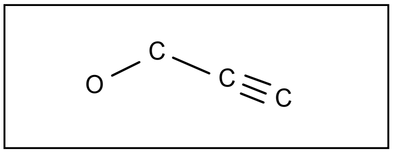


Figure 6-3: Molecular structure of 2-propyn-1-ol illustrating competing dipole bonds.

For multi-functional alcohols, all available methods estimated these components with extremely high deviations. Consider the estimation of hydrocarbon diols and triols with the previous best estimation of 16.79K (Table 6-6). With the inclusion of an OH-OH group interaction parameter, the deviation has been reduced to 8.43K. Subsequently, this was the starting point for the development of group interaction (Chapter 5). Thus, 13 OH and 7 OH(a) interaction parameters were developed. The results for both interaction groups are condensed and presented in Table 6-6. As a result, the inclusion of these group interactions has accompanied a far better estimation in the prediction of all alcohol compounds.

For multi-functional compounds, there were large deviations for cases with high dipole moments (Figure 6-4) and competing dipole bonds like 1,2-hexanediol (30.6K) and N,N- Bis-2-hydroxyethyl-piperazine (35K), respectively. These are the only 2 components with a deviation greater than 25K. Cordes and Rarey had 14 components with a deviation greater than 25K, which was the smallest number from all the available methods. Thus, the proposed method produced a more accurate distribution and a lower probability of extremely high deviations occurring.

Table 6-6: Functional analysis of alcohol compounds showing the deviations and number of components for the different models used.

	Number of Components						**Absolute Average Deviation (K)**					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
Mono-Functional Compounds												
All alcohols	150	150	148	150	131	150	**6.15**	25.57	6.57	12.35	13.24	6.77
1-alcohols	18	18	18	18	18	18	**8.32**	32.42	6.97	9.61	18.60	9.93
Primary alcohols	46	46	46	46	43	46	**6.53**	20.29	6.78	9.61	14.16	7.29
Secondary alcohols	45	45	44	45	44	45	**5.94**	25.54	5.32	11.54	13.19	5.85
Tertiary alcohols	31	31	31	31	18	31	**4.57**	39.97	6.76	19.88	17.90	6.65
Aromatic alcohols	32	32	31	32	30	32	**7.33**	18.87	8.26	9.43	8.49	7.99
Diols, Triols	22	22	22	22	22	22	**8.43**	32.05	26.78	28.12	18.79	16.79
Group Interactions												
OH	130	125	127	117	108	129	**8.45**	28.4	14.9	18.5	13.6	11.6
OH (a)	22	22	22	21	16	22	**9.60**	35.7	19.2	19.4	18.7	16.0
All Compounds												
All alcohols	337	331	331	320	263	336	**8.04**	26.78	11.60	15.66	13.96	9.98
Primary alcohols	142	139	139	134	122	141	**7.92**	19.27	13.85	14.92	13.59	10.56
Secondary alcohols	97	95	96	93	84	97	**7.99**	31.03	9.81	14.51	14.33	8.80
Tertiary alcohols	49	48	48	45	27	49	**6.30**	42.36	8.42	22.10	18.81	8.88
Aromatic alcohols	63	63	62	62	43	63	**10.37**	26.78	13.97	16.02	12.21	12.68
Diols, Triols	37	36	36	37	35	37	**9.76**	32.50	22.82	24.16	16.54	14.43

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

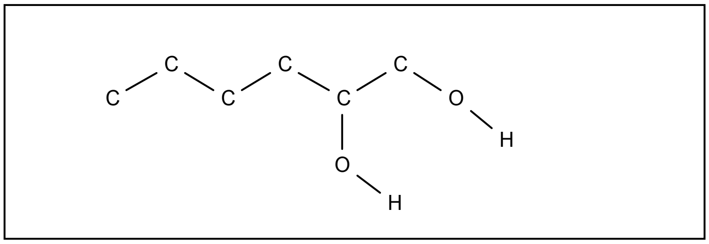


Figure 6-4: Molecular structure of 1,2-hexanediol illustrating high dipole moment.

6.2.2 Carbonyl Compounds

For the estimation of mono-functional carbonyl compounds, the proposed method yielded a more accurate estimation than previous methods, for all cases (Table 6-7). Although, Marrero and Pardillo reports a slightly lower deviation for ketones, the data set incorporated for ketones is much smaller, 41 as compared to 66. To establish this comparison, the 41 components were filtered and the proposed method revealed a deviation of 5.48K. The re-definition of aldehyde and ketone groups into separate non-aromatic and aromatic groups appears to have improved the estimations for these components. Other carbonyl compounds refer to non-cyclic carbonates, formic acid esters, lactones and acid chlorides, Table 5-3. Overall, the distribution for mono-functional compounds for the proposed method was good. In a few cases for carboxylic acids, large deviations occurred for long chain compounds, like 9-octadecenoic acid (24.5K) and abietic acid (17.3K). In general, carboxylic acids are the most complicated carbonyl group, since they are always found at the beginning or end of a chain, and the effect of dipole moment and hydrogen bonding is more effective.

The estimation of multi-function carbonyl compounds is probably more complex than alcohol compounds. For this reason, there were 4 interaction groups introduced (Table 6-7), which accounts for 36 group interaction parameters. The results, however, for the interaction groups are far more accurate, especially when reducing components with extremely high deviations. Thus the effect of the introduction of group interaction parameters can be observed by the better estimation and distribution of all carbonyl compounds. The correction introduced, C=C-C=O, also provided a better estimation of carbonyl corrections. The correction accounts for the oxygen atom inducing electrons not only from the carbon of a carbonyl group, but also from another sp^2 carbon in close proximity. Consequently, the larger charge separation produces a stronger dipole moment and polarisability. There were large deviations for components with strong steric effects like isophorone (26.4K) and 9-fluorenone (21.2K). For the latter component, there are 2 carbon-carbon π-bond in conjugation with the carbonyl double bond (Figure 6-5). Due to the limited database for these types of components, no special correction was introduced.

Table 6-7: Functional analysis of carbonyl compounds showing the deviations and number of components for the different models used.

	Number of Components						Absolute Average Deviation (K)					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
Mono-Functional Compounds												
Aldehydes	26	26	26	26	20	26	**5.52**	13.93	7.79	9.52	7.69	8.41
Ketones	66	60	65	57	41	66	**5.73**	13.63	8.55	8.07	5.68	7.36
Carboxylic acids	34	34	34	34	31	34	**6.51**	34.70	7.05	13.44	12.33	6.69
Esters	110	110	110	103	87	110	**4.26**	16.60	4.27	7.68	6.83	5.27
Other carbonyls	40	19	40	23	17	40	**2.79**	29.14	56.02	16.08	3.15	4.32
Group Interactions												
Aldehyde	19	19	17	17	11	19	**6.86**	26.59	15.02	14.30	12.80	14.83
Ketone	67	50	61	52	36	67	**9.11**	24.46	14.73	14.02	9.97	12.84
Carboxylic acids	16	16	16	15	15	16	**10.25**	29.07	12.46	34.55	12.32	12.60
Esters	152	146	147	136	110	152	**9.65**	34.60	12.78	13.30	12.75	11.64
Second-order Correction												
C=C-C=O	135	104	124	96	59	135	**6.06**	27.64	14.73	12.51	13.35	9.39
All Compounds												
Aldehydes	47	47	45	45	31	47	**6.48**	19.52	11.53	11.99	9.50	11.57
Ketones	138	114	131	112	78	138	**7.28**	19.58	11.92	10.85	7.69	10.52
Carboxylic acids	60	60	60	57	48	60	**8.57**	30.90	9.85	20.27	12.58	8.93
Esters	283	279	278	254	195	283	**7.70**	26.92	9.20	11.93	10.02	9.45
Other carbonyls	54	25	54	25	19	54	**5.01**	26.91	58.19	16.72	5.10	6.07

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

There were large deviations occurring for a number of multi-functional components involving group interaction parameters obtained from the Beilstein database. Despite the large deviations, these components were readily added to the data set, since experimental data was urgently needed for the regression of group interaction parameters. However, the reliability of the data is questionable. In most cases, the estimations were good, but in some cases the estimations were relatively high, like ethyl isopentyl malonic acid diethyl ester (31.3K). For DDB components, with the exception of carboxylic acids, there were no cases of extremely high deviations (> 25K) for carbonyl compounds. In the case of carboxylic acids, there were high deviations for components with relatively high dipole moments, like fluoroacetic acid (38.2K) and 2-methoxyphenylacetic acid (28.3K). For the first case, this component is the first in its series and incorporates a fluorine group, and thus should be considered too complex for a group contribution estimation. For the latter case, this component has two isomers, 3-methoxyphenylacetic acid (1.3K) and 4-methoxyphenylacetic acid (4.8K). These set of components exemplify the common problem associated with group contribution. For the first component, the dipole moment is extremely high, whereas the latter component has competing dipole bonds, although the electro-negativity for the acid is far greater than the ether group. Thus, group contribution cannot distinguish between these components, and will generally choose an optimized value to fit all components, in this case, the middle component.

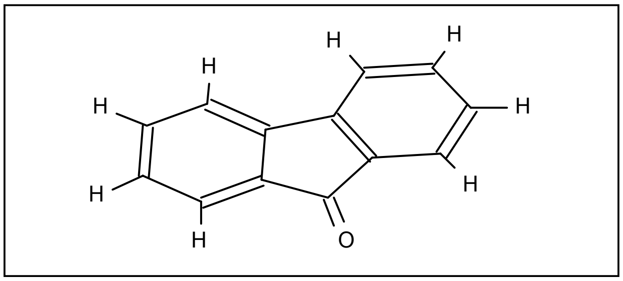


Figure 6-5: Molecular structure of 9-fluorenone illustrating larger charge separation.

6.2.3 Other Oxygenated compounds

With the exclusion of new groups, these compounds involve ethers, epoxides and anhydrides. In all cases, the estimations of these compounds yielded a far better result than previous methods (Table 6-8). The re-definition of the anhydride group into a chain and ring structure, again, establishes the development of the method on a functional basis. This allows these types of phenomena and differentiation to be identified. For the case of mono-functional compounds, there were no high deviations for these compounds (> 20K).

The development of group interaction led to the introduction of group interaction parameters for only ethers, aromatic oxygen and epoxides, due to the lack or property data for anhydride compounds. However, group interaction is based upon strong associating groups which attain non-additivity when the frequency is greater than one. For the case of ethers and epoxides, the groups are 'shielded' by two carbon atoms and even though the oxygen atom can induce electrons from both carbons, the effects of intermolecular interaction are hindered. Obviously this will also depend on the polarized charge on both carbon atoms. Nevertheless, these groups were introduced producing 14 and 4 interaction parameters for ethers and epoxides respectively. For the case of aromatic oxygen, even though the oxygen is shielded, it is connected to two aromatic carbons (sp^2). Thus, it can be considered that the oxygen atom has a much larger polarized charge than the cases above. Consequently, molecular interaction is more probable for this group, and 7 interaction parameters were generated. The above effect can relate to the results produced for these interaction parameters. For ethers and epoxides, the effect of molecular interaction is not as prominent as for aromatic oxygen.

The general difficulty with group interaction is the lack of property data for interaction parameters where only a single component exists. Consider the case of epoxides, there are 4 interaction parameters describing only 6 components, 3 of these parameters have only a single component.

The overall analysis of these compounds was, in all cases, far more accurate and had a better distribution. Large errors occurred for rather exotic molecules like, 1,1,3,3,5,5,7,7-Octaphenylcyclotetrasiloxane (Figure 6-6). With 4 interactions for the Ether-Ether interaction parameter, the proposed method estimated this structure with a deviation of 33.4K (3.5%), as compared to 51.0K by the previous method. Other methods could not estimate this structure. For the above component, the effect of steric hindrance is more influential than group interaction, considering also that silicon has a larger radius than carbon. Similar deviations occurred for these types of silicon ether components, but there did not seem to be a trend to befit an introduction of a correction (Table 6-9). For such large silicon ether structures, there could be a number of structural and physical effects occurring. There were also large deviations for components with extremely high dipole moments, like trichloromethyl ether (27.1K) and 2-methoxyphenylacetic acid (28.3K). These deviations are to be expected.

Table 6-8: Functional analysis of other oxygenated compounds showing the deviations and number for components of the different models used.

	Number of Components						Absolute Average Deviation (K)					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
Mono-Functional Compounds												
Ethers	98	97	98	78	77	98	**4.94**	9.90	6.61	8.93	5.22	5.78
Aromatic oxygen	5	5	0	0	2	5	**4.58**	22.18	-	-	8.61	7.36
Epoxides	12	12	12	10	9	12	**3.28**	12.07	9.61	28.29	5.62	7.12
Anhydrides	7	7	7	4	0	7	**8.33**	26.90	45.65	25.05	-	25.45
Group Interactions												
Ethers	303	264	271	210	210	302	**7.84**	17.49	11.78	15.78	9.96	8.99
Aromatic oxygen	13	10	0	0	8	13	**1.91**	14.48	-	-	6.51	13.18
Epoxides	6	6	6	6	5	6	**4.49**	33.63	26.86	39.61	12.19	10.55
All Compounds												
Ethers	458	400	411	312	297	455	**7.36**	16.45	12.04	14.00	8.82	8.34
Aromatic oxygen	18	15	0	0	10	18	**2.65**	17.05	-	-	6.93	11.57
Epoxides	20	20	20	18	14	20	**4.61**	22.02	16.00	33.47	7.96	8.22
Anhydrides	9	9	9	4	0	9	**9.17**	35.96	50.91	25.05	-	29.54

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

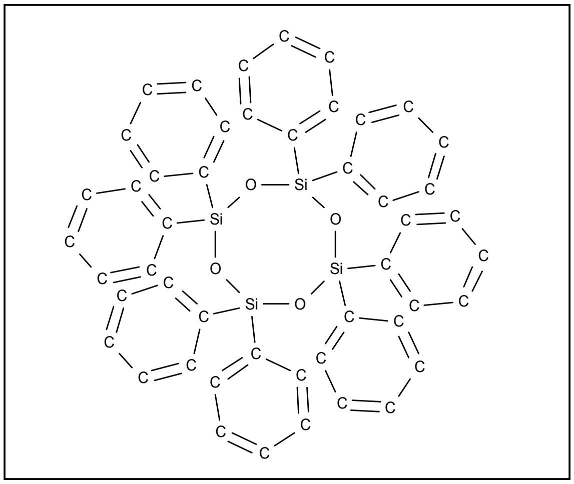


Figure 6-6: Molecular structure of 1,1,3,3,5,5,7,7-Octaphenylcyclotetrasiloxane.

The results for the introduction of new oxygenated structural groups are presented in

Table 6-10. The proposed method yielded a more accurate estimation for these compounds as compared to previous methods. It should be noted at this point that the structural definition of a group plays a major role in the estimation of a component. The structural group has a certain scientific definition and must be designed to fulfil this definition. For example, the proposed method defines an ester group such that, the C=O has only a carbon and oxygen attached, of which the latter has to be connected to a carbon atom. But if the carbon connected to the C=O is replaced by a nitrogen, because of the less electronegative nitrogen, the whole scenario changes and the ester group cannot capture this effect. Thus, in this case, the nitrogen ester group (carbamate) should not be fragmented as an ester group. Previous methods had such simple structural group definitions to allow the group to be fragmented as a carbonyl and amine group and this was the major reason for high deviations observed for these components. Overall, there were no relatively high deviations for these compounds (> 12K) for the proposed method. As with the case of anhydrides, because of the lack of property data, there were no group interaction parameters and only a single component for peroxides was found.

Table 6-9: Normal boiling points and deviations for silicon ether components.

Compounds	Boiling Point (K)	Deviation (K)
1,3,5-Trimethyl-1,1,3,5,5-pentaphenyltrisiloxane	754.0	-34.9
Dodecamethylpentasiloxane	532.9	34.6
1,1,3,3,5,5,7,7-Octaphenylcyclotetrasiloxane	961.2	33.4
Diphenyl-di-trimethylsiloxy-silane	579.0	-25.2
1,3-Bis-acetoxymethyl-1,1,3,3-tetramethyl disiloxane	483.7	-29.9

6.3 Nitrogen Compounds

6.3.1 Amine Compounds

The estimation for mono-functional amine compounds yielded a far more accurate estimation than previous methods (Table 6-11). The only exception is the Marrero and Pardillo method for the estimation of primary amines. However, the estimation was based upon 6 less components. Filtering these components revealed a deviation of

4.81K for the proposed method. An overall analysis of primary amines revealed no high deviations (> 15K). For the case of secondary and tertiary amines, the development and re-definition of structural groups to distinguish between different structural behaviours allowed a more accurate estimation. There were large deviations for the first components in their respective series, trimethyleneimine (19.2K) and trimethylamine (18.6K). Apart from these components, there were relatively no high deviations for these types of components. Consequently, the proposed method yields an excellent distribution for the estimation of amine compounds.

Table 6-10: Deviations and number of components of models for new structural groups involving oxygenated compounds.

	Number of Components						**Absolute Average Deviation (K)**					
Compounds	**Pr**[a]	**JR**[b]	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
						Mono-Functional Compounds						
Cyclic carbonates	2	2	2	0	0	0	**2.89**	96.44	62.33	-	-	-
Peroxide	1	0	0	0	0	0	**0.00**	-	-	-	-	-
Carbamates	6	4	6	5	0	0	**4.90**	20.19	48.58	28.06	-	-
						All Compounds						
Carbonyl with S	4	4	4	2	0	0	**3.91**	27.64	64.83	42.72	-	-
Carbamates	11	4	8	6	0	0	**6.13**	20.19	47.95	25.80	-	-

[a] Proposed method, [b] JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey

Since tertiary amines are 'shielded' by three carbon atoms, the effect of group interaction is extremely small. The effect of steric hindrance is probably more significant in this case, such that the electronic system is hardly available for any interaction with the nitrogen. An analysis of tertiary amine compounds with interaction groups also revealed no need in the introduction of amine as an interaction group. Thus, only primary and secondary amines were chosen as interaction groups, which produced 10 and 7 interaction parameters, respectively. These parameters probably produced the best results from all interactions. Consider that the estimation of multi-functional amine compounds is just as accurate as mono-functional compounds. There were no extremely high deviations for these compounds. This verifies the exceptional distribution produced by the proposed method. The only exceptions were, 3,4-dichloro aniline (22.8K), N,N-dimethyl-3-nitroaniline (21.6K), 1,4-dimethyl piperazine (22.1K) and N,N-Bis-2-hydroxyethyl-piperazine (35K). The latter four components are tertiary amines and the effect of steric hindrance and dipole moment is in all probability more significant. The overall analysis of all compounds revealed a far better estimation than previous methods, with the exception for tertiary amines for the Marrero and Pardillo method. However in their case, the better estimation is obtained as a result of employing 26 components as compared to 56 for the proposed method (7.89K for the 26 components).

Table 6-11: Functional analysis of amine compounds showing the deviations and number of components for the different models used.

	Number of Components						**Absolute Average Deviation (K)**					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
Mono-Functional Compounds												
Primary Amines	43	43	42	41	37	43	**5.06**	20.23	7.22	13.14	4.66	7.94
Secondary Amines	36	36	36	35	30	36	**4.57**	14.06	10.35	12.37	5.65	8.86
Tertiary Amines	18	18	18	18	16	18	**7.05**	17.46	10.68	9.08	8.00	7.95
Group Interactions												
Primary Amines	44	43	42	39	36	44	**5.24**	16.67	13.92	15.18	10.29	10.76
Secondary Amines	23	21	21	19	15	23	**3.24**	19.41	17.05	17.30	9.75	7.54
All Compounds												
Primary Amines	93	92	90	86	75	93	**5.44**	18.23	10.61	14.56	7.93	9.14
Secondary Amines	65	59	60	57	45	63	**4.38**	17.20	14.05	13.97	7.02	8.48
Tertiary Amines	56	37	42	39	26	55	**9.25**	20.10	23.18	18.74	9.08	10.99

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

6.3.2 Amide Compounds

Table 6-12 provides a comparison for the different types of amide compounds. In all cases, the proposed method yielded the lowest deviation. It should be noted at this point that, although there has been no development in amide compounds, the better estimation is attained by the stronger differentiation in all structural groups. Second-order corrections and interactions negate certain effects that first-order groups are unable to capture. Consequently, the estimations are favourable for all compounds within the range of the method. For the case of amides, there were large deviations for the first components in its series, like acetamide (12.4K), N-1,1-dimethylethyl formamide (18.5K) and diethylcarbamic chloride (19.6K). The latter component also has an extremely high dipole moment and polarizability due to the inclusion of a chlorine atom. These results substantiate the inability of group contribution methods to accurately estimate the first components in a chemical series and for cases of extremes in the dipole moment and polarizability.

Table 6-12: Functional analysis of amide compounds showing the deviations and number of components for the different models used.

	Number of Components						**Absolute Average Deviation (K)**					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
Mono-Functional Compounds												
Amides	4	3	4	3	0	4	**7.43**	111.0	47.77	8.04	-	8.58
Mono Amides	6	4	6	0	0	6	**7.53**	83.59	22.77	-	-	12.51
Di-Amides	8	2	8	5	0	8	**5.64**	74.45	11.38	19.16	-	6.80
All Compounds												
Di-Amides	10	3	9	5	0	10	**8.87**	63.44	12.26	19.16	-	9.31

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

6.3.3 Other Nitrogen Compounds

The results for these compounds, with the exception of new groups, are presented in Table 6-13. For the estimation of mono-functional compounds, deviations for the proposed method are among the lowest achieved. A comparison with Marrero and Pardillo and Constantinou and Gani in certain cases could not be undertaken due to these methods employing much smaller data sets. There were only large deviations for the case of nitromethane (22.2K), which is the first compound in its series, and 3,4-dimethylpyridine (24.5K). For the latter compound, it has a number of isomers with very different dipole moments. In addition, this compound includes a correction from the previous method which involves ortho pairs on an aromatic ring. For many components, it was found that this correction should not be included in aromatic nitrogen rings. In future, this discovery will need to be corrected. For the case of oximes, the slightly higher deviation is as a result of the different dipole moments for the components. For oximes, the components generally have a higher dipole moment when the carbon part of the group has only one carbon neighbour, which would imply that the group is at the beginning or end of the molecular chain. The analysis of these compounds revealed that 3 out of the 9 components had a much lower dipole moment as a result of the carbon having two neighbours. Consequently, these components produced slightly higher deviations (10 to 15K).

For the development of group interactions, 5 and 6-membered nitrogen rings, cyanides, isocyanates and nitro groups were chosen as interaction groups, which produced 2, 7, 6, 1 and 6 interaction parameters respectively. It can be considered that these are strong associating groups and the effect of intermolecular interaction is much stronger. Consequently, the interaction parameters generated for these types of compounds, accounted for the reduction in the higher deviation multi-functional compounds. There was only a high deviation for nitrotrichloromethane (28.9K), which is the first component in its series and has three chlorine atoms. Consequently, this component can be considered too complex for a group contribution estimation. Overall, the proposed method produced an excellent distribution for these set of compounds.

The reliability of experimental data is an extremely important issue, as discussed in Chapter 5. Components are not omitted because of high deviations, but rather on their reliability. Consider the case of 2-nitrophenyl isocyanate (462.4K) and its isomers, 3-nitrophenyl isocyanate (460.0K) and 4-nitrophenyl isocyanate (460.1K). The estimation of these compounds, without group interactions, produced an average deviation greater than 40K. With an isocyanate and nitro interaction parameter, this deviation was reduced to less than 3K. However, these components, which were obtained from the same source, cannot be included because of their reliability. Firstly, consider that these isomer components have completely different dipole moments, yet, the experimental boiling point for all components is within 2.5K. Secondly, a vapour pressure measurement was obtained from Beilstein, which revealed a value of 435K at 20mmHg. Hypothetically, if it is possible to extrapolate this measurement to atmospheric pressure (cannot be done as there is only 1 point), a value of 460K is unrealistic. Consequently, the prediction capability of the method would now be highly questionable.

Table 6-13: Functional analysis of other nitrogen compounds showing the deviations and number for components of the different models used.

	Number of Components						**Absolute Average Deviation (K)**					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
Mono-Functional Compounds												
5-membered N ring	12	9	12	0	1	12	**5.39**	44.52	29.27	-	0.11	7.14
6-membered N ring	28	28	28	20	14	28	**7.48**	13.20	7.52	21.60	4.13	8.99
Cyanides	26	26	26	19	23	26	**4.82**	22.42	3.38	7.91	8.45	10.09
Isocyanates	9	0	9	0	0	9	**5.44**	-	69.94	-	-	9.90
Oximes	9	0	0	0	0	9	**8.00**	-	-	-	-	7.88
Nitrites	15	8	8	7	1	15	**5.26**	28.81	12.37	6.03	0.01	7.82
Nitrates	5	0	0	0	0	5	**2.95**	-	-	-	-	3.83
Group Interactions												
5-membered N ring	11	9	0	0	2	11	**8.35**	15.45	-	-	6.90	18.50
6-membered N ring	14	12	14	7	9	14	**6.53**	26.05	14.87	28.02	17.05	16.51
Cyanides	15	15	14	5	9	15	**7.06**	34.34	18.26	14.81	13.95	17.06
Isocyanates	3	0	3	0	0	3	**4.33**	-	94.84	-	-	7.88
Nitro	13	13	13	13	0	13	**4.49**	73.59	18.24	16.05		21.74
All Compounds												
5-membered N ring	23	18	12	0	3	23	**6.81**	29.98	29.27	-	4.64	12.57
6-membered N ring	44	41	44	28	24	44	**7.03**	17.95	9.57	23.38	8.89	11.32
Cyanides	44	44	43	24	33	44	**5.73**	27.94	8.37	9.35	9.92	12.96
Isocyanates	16	0	15	0	0	16	**5.34**	-	74.33	-	-	8.92
Nitrites	42	34	35	31	1	42	**6.61**	57.06	15.07	13.21	0.01	13.59
Nitrates	6	0	0	0	0	6	**3.48**	-	-	-	-	4.30

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

The results for the estimation of new structural groups developed for nitrogen compounds are presented in Table 6-14. In all cases, the proposed method yields the lowest deviation. There were also no high deviations for these components (> 15K), thus the distribution is exceptional. Due to the lack of property data, these groups were not introduced as interaction groups. For the case of imidazole, all 4 components had a frequency of 2 for this group. Subsequently the group also functions as an interaction parameter.

Table 6-14 Functional analysis of new structural groups involving nitrogen compounds showing the deviations and number of components for the different models used.

	Number of Components						Absolute Average Deviation (K)					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
						Mono-Functional Compounds						
Urea	1	1	1	1	0	0	**0.00**	56.90	5.44	35.58	-	-
Imidazole	4	0	4	0	0	0	**9.34**	-	62.53	-	-	-
Quaternary amine	3	0	0	0	0	0	**5.34**	-	-	-	-	-
(C=) amine	3	3	3	0	0	0	**4.45**	53.37	10.03	-	-	-
Isoazole	3	0	0	0	0	0	**5.05**	-	-	-	-	-
						All Compounds						
(C=) amine	6	6	6	1	0	0	**5.90**	55.26	14.98	21.19	-	-
Cyanamides	1	1	0	0	0	0	**0.00**	51.76	-	-	-	-
Thiocyanates	3	3	0	0	0	0	**6.51**	22.55	-	-	-	-

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

6.4 Sulphur Compounds

The results of the estimations for sulphur compounds are presented in Table 6-15. In all cases for mono-functional compounds, the proposed method yielded the lowest deviation. The only exception is the Marrero and Pardillo method, which employs a lesser number of components. It should be noted that, as discussed in Chapter 2, the range of the Marrero and Pardillo method is fairly limited. A large number of components cannot be fragmented by this method, thus the predictive capability is problematic. Overall, there was only a large deviation for thiacyclobutane (16.2K), a cyclic structure which is the first component in its series.

Thiol, thioether and aromatic thioether were chosen as interaction groups which generated 3, 4 and 4 interaction parameters respectively. As before, lack of property data prevented introduction of other sulphur groups as interaction groups. The introduction of these sulphur interaction parameters yielded a good estimation of multi-functional compounds. There were large deviations for compounds involving a large number of halogens on a group, thus affecting the polarizability and dipole moment, like diperfluoromethylthioether (18.6K) and 2,5-bis-trichlorosilyl thiophene (28.7K). For the latter component, the estimation is also influenced by the steric hindrance due to 2 silicon groups, each having three chlorines. Apart from these components, the proposed method produced an outstanding distribution with relatively few or no high deviations.

Table 6-15: Functional analysis of sulphur compounds showing the deviations and number of components for the different models used.

	Number of Components						**Absolute Average Deviation (K)**					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
							Mono-Functional Compounds					
Disulfides	4	4	0	4	3	4	**2.97**	7.96	-	20.17	1.12	3.38
Thiols	37	37	37	18	18	37	**3.64**	12.91	7.21	7.60	7.46	5.13
Thioether	30	30	30	28	19	30	**4.86**	15.45	8.05	11.49	3.88	4.56
Aromatic Thioether	10	10	0	8	3	10	**3.68**	11.04		8.68	0.25	5.17
Sulfolane	3	0	0	0	0	3	**4.47**	-	-	-	-	12.11
Isothiocyanates	3	0	0	0	0	3	**2.08**	-	-	-	-	8.61
							Group Interactions					
Thiol	12	12	12	12	12	12	**3.62**	32.34	8.40	30.39	21.56	6.82
Thioether	17	17	17	11	7	15	**6.00**	20.82	20.22	19.68	7.48	12.96
Aromatic Thioether	15	10	0	4	2	15	**6.58**	18.87	-	21.04	6.90	17.11
							All Compounds					
Thiols	50	50	50	30	30	50	**3.79**	18.21	8.38	16.72	13.10	5.58
Thioether	56	56	53	44	26	48	**5.58**	19.18	14.64	14.84	4.85	7.98
Aromatic Thioether	31	23	0	12	5	31	**6.83**	14.24	-	12.80	2.91	12.07

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

The results for the estimation of new structural groups for sulphur compounds are presented in Table 6-16. The proposed method is the only method able to predict these types of compounds. The estimations of these compounds are extremely good; however, the estimation is based upon a small data set. These types of complicated sulphur groups are in general, less stable. The poor stability is as a result of the strong associating sulphur groups existing in the molecule. They do not have experimental normal boiling points because they are non-existent. In other words, for cases of longer chains and multi-functional compounds involving another strong associating group, the components decompose before the boiling point is reached. For example, a sugar molecular is a poly-glycol incorporating a hydrocarbon chain with 11 carbon atoms. Due to the large number of alcohol groups, the molecule decomposes before the normal boiling point is reached. This is the general reason why there is a limited range of experimental data for these types of components. For critical properties, the experimental data set is even smaller (15 to 25% of components that have normal boiling points).

Table 6-16: Functional analysis of new structural groups involving sulphur compounds showing the deviations and number for components of the different models used.

	Number of Components						Absolute Average Deviation (K)					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
Mono-Functional Compounds												
Sulfates	2	0	0	0	0	0	**2.07**	-	-	-	-	-
Sulfon amides	3	0	0	0	0	0	**6.52**	-	-	-	-	-
Sulfoxide	1	0	0	0	0	0	**7.74**	-	-	-	-	-
All Compounds												
Sulfoxides	2	0	0	0	0	0	**7.49**	-	-	-	-	-

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

6.5 Halogen Compounds

The results for the estimation of halogenated compounds are presented in Table 6-17. The proposed method yielded the lowest deviation for all the different types of halogen compounds. The only exception is the previous method for the case of saturated chlorinated compounds, which produces a slightly lower deviation. In general, the estimation of saturated halogenated compounds should yield a better estimation than unsaturated compounds. With unsaturated compounds involving double or triple bond carbons attached to the halogen, the ability of the electronegative halogen to polarize the unsaturated carbon produces a larger charge separation than the case with a saturated carbon. For the case of fluorine and bromine, the results do testify to this effect. For chlorine, although the above effect is prominent, the results do not verify this. Consider that in general, a functional group has a contribution value dependant on the range of components. For example, a fluorine group will always have a factor for the dipole moment in the contribution since the component it represents generally has a dipole moment. An analysis of chlorine groups revealed a large number of components that included three chlorines connected to a carbon atom. This represents one of the extremes for the dipole moment, and even with the introduction of a C-[F,Cl]$_3$ correction which generally tends to reduce the deviation, an accurate estimation is not possible by a group contribution method. The correction was able to provide a better estimation overall, but in some cases there are other electronegative groups in the molecule which also affect the dipole moment, for example, 2,2,2-trichloromethyl ether (27.1K). The same applies to the other extreme, which led to the introduction of another correction (C)$_2$-C-[F,Cl]$_2$.

Since one of the major aims of the proposed method was to reduce components with particularly high estimations; the corrections described above are feasible. Only in a few cases, have the corrections slightly increased the deviations. These are for cases of compounds which do not behave as expected, which was the major reason for the introduction of the corrections. An analysis of these compounds yielded a large number of smaller components incorporating other electronegative functional groups. Reflect on the estimation of a smaller and larger molecule, both having the same dipole moment. Apart from the general difficulty associated with the first components in a series, the smaller compounds also have a greater kinetic energy as a result of their smaller masses. Thus, the tendency to escape into the vapour phase as a result of the greater kinetic energy is far greater than for a larger molecule. Even with the introduction of corrections, which attempts to capture the two extremes, these types of compounds should be considered exotic. Typical cases are tribromoacetaldehyde (30.2K), 2,2,2-trichloroethanol (24.3K), 2,2,2-trifluoroethanol (27.6K), nitrotrichloromethane(28.9K), 1,2,2-trichloropropane (19.5K) and 5-(2-chloro-1,2,2-trifluroethoxy)-1,1,2,2,3,3,4,4-octafluoropentane (29.8K). In these cases, the corrections are conflicted by other highly electronegative groups and also the size of the molecule.

Table 6-17: Functional analysis of halogen compounds showing the deviations and number of components for the different models used.

	Number of Components						**Absolute Average Deviation (K)**					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
						Mono-Functional Compounds						
Saturated fluorinated	64	64	64	64	63	64	**6.38**	18.53	59.89	25.62	14.25	7.35
Fluorinated	86	86	86	86	83	86	**7.17**	18.01	48.96	26.46	12.83	8.02
Saturated chlorinated	64	64	64	60	0	64	**7.53**	26.01	14.72	8.26	-	7.43
Chlorinated	117	117	117	95	0	117	**6.96**	21.36	14.05	9.36	-	6.62
Saturated brominated	31	31	31	31	0	31	**5.48**	14.76	8.82	6.91	-	6.34
Brominated	49	49	49	49	0	49	**5.83**	12.89	8.88	7.48	-	7.36
Iodinated	18	18	18	18	15	18	**5.13**	13.63	6.91	5.77	12.06	5.20
						Second-order Corrections						
$(C=O)-C([F,Cl]_{2,3})$	19	19	19	15	2	19	**12.39**	34.41	28.23	30.45	3.62	20.14
$(C=O)-(C([F,Cl]_{2,3}))_2$	2	2	2	0	1	2	**0.43**	66.87	34.30		7.78	47.30
$C-[F,Cl]_3$	139	136	137	111	52	138	**8.33**	26.56	43.06	33.46	11.48	10.94
$(C)_2-C-[F,Cl]_2$	69	69	69	57	45	69	**8.41**	21.73	70.34	26.81	14.18	9.39
						All Compounds						
Fluorinated	213	203	205	169	116	212	**8.13**	24.55	38.16	32.55	12.86	10.10
Chlorinated	308	267	270	213	0	308	**8.18**	24.34	15.55	15.12	-	8.40
Brominated	94	92	93	90	0	94	**7.57**	17.97	13.67	15.73	-	8.69
Iodinated	28	28	28	27	20	28	**5.82**	18.09	9.94	12.48	12.39	7.03

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

The complexity of halogenated compounds can be clearly observed by the number of corrections included. These include 4 corrections, 2 of which have been already discussed. The other 2 include carbonyl halogen corrections which also attempt to account for the dipole moment, in this case, with carbonyl halogen compounds. The design of these corrections is based upon a set of components which produce high deviations. If these corrections were not introduced, the subsequent high deviations will affect the group contribution and the predictive capability. For example, for the carbonyl halogen corrections, this correction removes the effect of extreme polarisability and dipole moment. If this correction was not introduced, then the large deviations observed will affect the first-order group contribution values and subsequently the prediction of all compounds.

Overall, the estimation of halogenated compounds will always be a problem for group contribution methods, since they occur more frequently than other electronegative groups. Consider for example, the estimation of iodinated compounds. For these compounds, the group only appears once, with only one exception where it appears twice. In this case, the positioning is not as significant and since it does not occur as frequently as other halogens, the estimation produces a far lower deviation. Consequently, the influence on the polarisability and dipole moment of a molecule is entirely based upon the positions of these groups and group contribution cannot capture this effect. Since these groups are also not as electronegative as other functional groups, the estimations are not as extreme. This proposes a good distribution but with a slightly higher average deviation. There were high deviations for chloroform (22.7K) 1,1-difluoroethane (20.1K) fluoroacetic acid (38.2K), ethyl triflourosilane (24.7K) diethylcarbamic chloride (29.6K) 3,4-dichloro aniline (22.8K) ethyl-2-chloro-propionate (43.6K) and 2-bromophenol (28.6K). These components are the first in their series coupled with other highly electronegative groups.

6.6 Other Elemental Compounds

The estimation of the various other elemental compounds as well as the new groups introduced is presented in Table 6-18. The proposed method yielded an extremely low deviation for these set of components. With the new groups, the proposed method also suggests the broadest range of applicability from all methods. For the case of silicon, the groups involving electronegative elements produced a higher mean deviation. Since carbon has similar characteristics as silicon, functional groups incorporated silicon as a possible neighbour, for example, chlorine connected to a carbon or silicon atom. Although, this maybe accurate for a large number of cases, there are certain cases involving stronger steric effects and a weaker electronegative potential due to the larger molecular weight and radius of silicon as compared to carbon. These cases produce larger deviations which were described previously (Section 6.2.3).

Table 6-18: Functional analysis of other elemental compounds showing the deviations and number for components of the different models used.

	Number of Components						Absolute Average Deviation (K)					
Compounds	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**	**Pr**	**JR**	**SB**	**GC**	**MP**	**CR**
Mono-Functional Compounds												
Phosphates	4	0	0	0	0	4	**4.97**	-	-	-	-	6.98
Arsine	6	0	0	0	0	6	**3.17**	-	-	-	-	3.07
Germanium	1	0	0	0	0	1	**0.00**	-	-	-	-	15.81
Germanium & Cl_3	3	0	0	0	0	3	**1.20**	-	-	-	-	6.03
Stannium	3	0	3	0	0	3	**1.14**	-	2.42	-	-	1.30
Borates	8	0	0	0	0	8	**6.35**	-	-	-	-	5.05
Silicon	37	0	27	0	0	37	**5.01**	-	19.82	-	-	5.06
Silicon to O	43	0	0	0	0	40	**10.77**	-	-	-	-	10.80
Silicon to F or Cl	80	0	0	0	0	77	**9.67**	-	-	-	-	10.12
New Groups												
Phosphine	4	0	4	0	0	0	**1.65**	-	14.72	-	-	-
Selenium	1	0	1	0	0	0	**0.00**	-	10.76	-	-	-
Aluminum	2	0	0	0	0	0	**5.50**	-	-	-	-	-

(Proposed method, JR – Joback and Reid, SB – Stein and Brown, GC – Constantinou and Gani, MP – Marrero and Pardillo, CR – Cordes and Rarey)

The general argument with the estimation of metal compounds would be the predictive capability of these groups, due to the smaller number of components used. This smaller set generally includes mono-functional compounds and the argument would be based on predictive capability of multi-functional compounds, in particular highly electronegative groups or anions. Since metal groups can act as cations, and with the case of multi-functional compounds involving anions, these compounds are now called ionic liquids. For these set of compounds, there is no vapour pressure. Thus, the predictive capability of these compounds can be considered good.

6.7 Model Development

The development of the normal boiling point model involved the analysis of the different models proposed in Chapter 5. The results of these models (Section 5.7) are presented in Table 6-19. The analysis was performed on a set of 2557 components excluding the Beilstein data set. For the regression, the criterion for convergence employed was $1\text{x}10^{-8}$.

Table 6-19: Average absolute deviation for the different models proposed (Section 5.7)

Equation no.	Average Absolute Error (K)	Equation no.	Average Absolute Error (K)
5-1	15.5126	5-2	6.6846
5-3	6.9714	5-4	6.6747
5-5	6.6723	5-6	6.7572
5-7	6.6756	5-8	8.7249
5-9	7.0407	5-10	7.0403
5-11	9.8546	5-12	6.6749
5-13	6.9567	5-14	6.9551
5-15	6.6748	5-16	6.6759
5-17	6.6844	5-18	6.6896

The first analysis involving fitting a logarithmic model produced the highest average deviation from all models (Equation 5-1). The previous method incorporated a model (Equation 5-2) which gives a good description of the dependence of the normal boiling point on molecular size. The model also produced one of the lowest deviations with only three non-linear parameters. The same model was then tested with the molecular weight instead of the number of atoms (Equation 5-3) and produced a slightly higher average deviation. This conclusively proves that number of atoms has a stronger influence than molecular weight on group contribution. The model from the previous method was also tested using the molecular weight in three different forms (Equations 5-4, 5-5 and 5-6). In all cases, these models did not produce a significant improvement (only 0.18% improvement for Equation 5-5) with the inclusion of another physical contribution and a large number of non-linear parameters. Consequently, employing the molecular weight with the number of atoms did not suggest a more meaningful result. The model was also tested using a power and logarithmic fit (Equations 5-7 and 5-8). For Equation 5-7, the value of the exponent was extremely close to one. Thus, in this case the average deviation is quite similar to the previous model. For the logarithmic fit, this produced a much higher average deviation. The higher average deviation is a result of the model not being able to predict higher temperature compounds (consider the fit of Constantinou and Gani and Marrero and Gani in Figure 6-1, both methods employing a logarithmic fit). Consequently, the logarithmic fit should not be considered in the model development of group contribution. It was discussed in Chapter 3 (Equation 3-11) that the number of atoms and molecular weight has a linear relationship to the normal boiling point. The testing of these models (Equations 5-9 and 5-10), however, produced a higher mean deviation. The model proposed by Retzekas et al (2002) involved a separate group and physical contribution (Equation 5-11). In this case, theresult was poor because of the competing contributions. The model involving the molecular weight as part of the numerator (Equation 5-12) also produced a similar deviation to the previous model. In other words, the inclusion of the molecular weight as a linear relationship to the normal boiling point did not show any improvement.

The previous model was also tested using various mathematical forms, including a quadratic fit (Equations 5-13 to 5-18). In all cases, there were no significant improvements over the original model. The development of a second set of contributions also produced higher average deviations. However, the regression for these contributions is quite complicated, since there are three different types of regression viz. non-linear, linear and successive approximation. Consequently, the starting values for the simplex algorithm were a major influence on the regression. This now plays a major role when the new simplex is formed. Thus, there were two different types of regression performed. The first type involved leaving the second set of contributions unchanged when a new simplex was built. The second type involved returning the original values when a new simplex was built. Both types were applied to the equations described in Chapter 5. For the case of fitting the second set of contributions instead of the number of atoms, this produced negative contributions for a few groups. Other cases involved, were based on fitting the contributions to the exponent of the number of atoms and summation of group contributions. However, this also produced higher deviations which can be attributed to the sensitivity of the exponent values.

For all the models tested, the previous model is probably the most feasible. The model only incorporates three non-linear parameters and a readily available quantity viz. the number of atoms. The model produces among the lowest average deviations and by the relationship of the experimental and calculated normal boiling points (Figure 6-7), an exceptional distribution. The relationship provides hardly any large outliers and is independent of the range of temperatures. Consequently, the model will be used for the development of the proposed method. All the results provided in this chapter, are based upon this model.

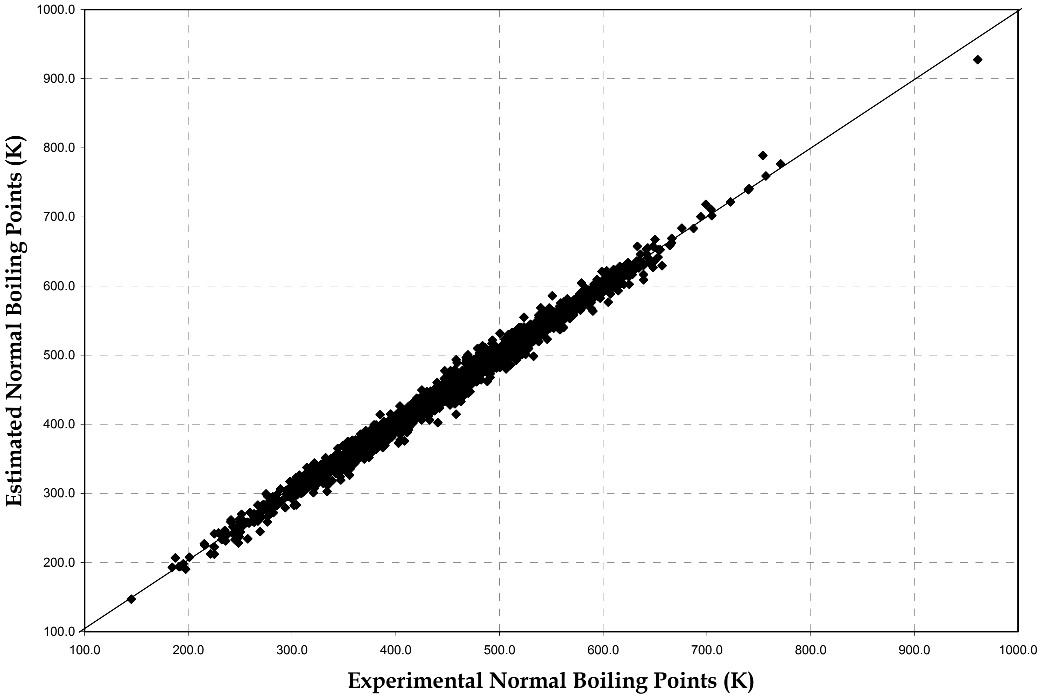


Figure 6-7: Relationship of the experimental and calculated normal boiling points for the proposed method.

6.8 Overall analysis

In order to test the predictive capability of the method, a data set was chosen from the Beilstein database which was not used in the regression. The data set comprised of 405 components common to all methods with the exception of the Marrero and Pardillo method (212 components in this case). For these sets of components, the proposed method yielded an average absolute deviation of 4.68K (19.04K for the Joback and Reid method, 7.67K for Stein and Brown, 12.09K for Marrero and Gani, 10.74 for Marrero and Pardillo and 6.30K for Cordes and Rarey). Since components were urgently needed for the model development of group interactions and functional groups with only a few components, these types of multi-functional components are not present in the test set. Also, the test set included only a few components with the correction C=C-C=O. No other corrections for the proposed method were in the test set. Thus, the proposed method yielded the most accurate estimation for these types of components, even though most of the second-order groups were not present.

Overall, the proposed method yielded an average absolute deviation of 6.50K (1.52%) for a set of 2820 components. For the available methods, Joback and Reid produced an average absolute deviation of 21.37K (4.67%) for a set of 2514 components, 14.46K (3.53%) for 2579 components for Stein and Brown, 13.22K (3.15%) for 2267 components for Constantinou and Gani, 10.23 (2.33%) for 1675 components for Marrero and Pardillo and 8.18K (1.90%) for 2766 components for Cordes and Rarey. This implies that the proposed method yielded the lowest average deviation with the broadest range of applicability.

The most important criterion for the reliability of a group contribution method is the probability of prediction failure. This involved the extreme deviations between the estimated and experimental normal boiling points. This relationship can be represented by Figure 6-8, which presents a plot of the fraction of data greater than a given temperature. The calculations were based on a common set of data for all methods compromising of 2177 components with the exception of Marrero and Pardillo (1546 components in this case). It can be clearly seen that the proposed method yields a far better distribution and lower probability of prediction failure. Consider for example, for the proposed method 3% of the data is greater than 20K, 6% for the

Cordes and Rarey method. The other methods range from 36% for Joback and Reid, 16% for Stein and Brown, 20% for Constantinou and Gani and 16% for Marrero and Pardillo.

The results presented in this chapter have proven all the objectives set out for the development of a group contribution method for the prediction of normal boiling points. In particular, the estimation of polycyclic multi-functional compounds has proven to be quite successful. Also, the proposed method is now able to differentiate between isomers. The success of the proposed method can actually be attributed to its identifiable weakpoints. Since group contribution has certain structural, physical and electronic limitations, estimation for these types of components will always be a problem, owing to the fact that the only required input is the molecular structure. With the procedure provided in this work, these limitations have already been identified, which in particular involves components with extremes in the dipole moment and the first few components in a series. The higher deviations provided in Figure 6-8 generally involve these types of components. For a more sophisticated estimation, the dipole moment can be obtained from a molecular mechanics calculation, which will provide a solution to the weakpoints mentioned in this chapter. For the first few components in a series, these data are readily available. However, this problem can be captured by the inclusion of a molecular property like molecular surface area. In addition, this molecular property can also capture steric effects and isomer differentiation in more detail. But in order to develop this type of method, the group contribution method has to be at its ‘full capacity’, whereby its limitations owe to those cases described above. The proposed method does exactly this. Thus, the reliability is now even more prominent with the expectation that, within the limitations of group contribution, estimations can be performed with confidence.

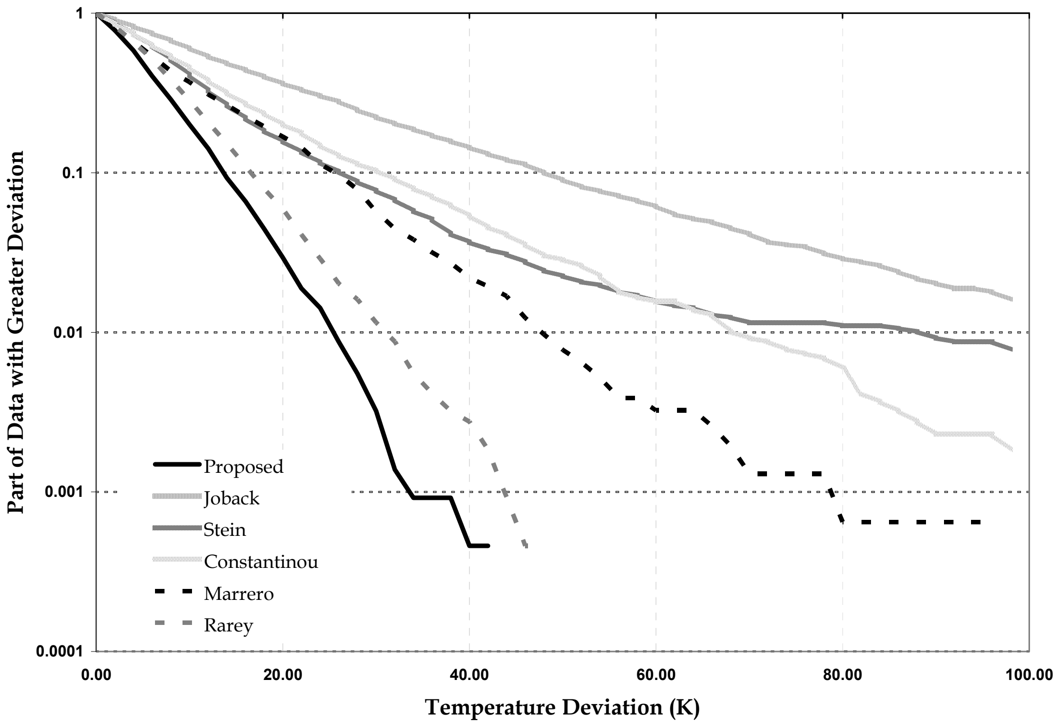


Figure 6-8: Part of data with deviations greater than a given temperature.

Chapter Seven

Conclusion

A group contribution method has been developed for the estimation of normal boiling point temperature for non-electrolyte organic compounds, which extended the work of Cordes and Rarey (2002). Group contribution is the simplest form of estimation and requires as input, only the molecular structure of the compound. Structural groups were defined in a standardized form and fragmentation of the molecular structures was performed by an automatic procedure to eliminate any arbitrary assumptions.

Owing to the large number of compounds (2820 components) this work suggested an analysis of groups on a functional basis. Consequently, structural groups and components belonging to a specific functional group were analysed. This allowed for the examination of the different types of phenomena or behaviour occurring within an organic molecule. For example, this analysis led to the re-definition of an anhydride group into a chain and cyclic group. This analysis was essentially accomplished by the implementation of a metalanguage filter program.

The structural first-order groups were defined according to its neighbouring atoms. This definition suggested a rather more scientific characterization of structural groups since it provided knowledge of the neighbourhood and electronic structure of the group. In this manner there were 115 first-order groups defined, which also provided the broadest range of non-electrolyte organic compounds as compared to current methods.

Second-order corrections were defined to those limited cases, in which larger structures, physical, electronic or structural effects could not be defined as structural groups. The corrections proposed by Cordes and Rarey (2002) were implemented in this work. Steric and isomer corrections were introduced to account for the steric hindrance within a molecule. The correction also enables a more accurate differentiation between isomers. There were also carbonyl and halogen corrections which, in particular, account for certain electronic effects which could not be captured by the structural groups. The major development, however, was the introduction of group interactions. These groups are designed for multi-functional components with more than one strongly associating structural group. For these types of components, the assumption of simple additivity is no longer observed.

Many different models relating the normal boiling point to group contribution were implemented. This involved different mathematical forms of this relationship, which also included models proposed by previous methods. From the testing of these models, the model proposed by Cordes and Rarey (2002) proved to be the most accurate.

Overall, the proposed method proved to be the most accurate group contribution method as compared to previous methods, and with the broadest range of applicability. The method is now able to predict mono-functional compounds for all functional groups with an exceptional low deviation. In particular, the inclusion of group interactions led to a more accurate estimation of multi-functional compounds. For these types of compounds, previous methods have produced drastically high deviations.

The reliability of the proposed model is quite good with relatively few cases of components with extremely high deviations noted. These cases involved the first few components in their respective series and components with only a single functional group. Since these components are widely available, no structural group or correction was introduced. There were cases of high deviations for components with an extremely high dipole moment. These types of components cannot be captured by group contribution, since the position of the electronegative group (s) in the molecule determine the dipole moment.

On a test set of 405 components common to all methods, except for the Marrero and Pardillo method (212 components in this case), the proposed method yielded an average absolute deviation of 4.68K (19.04K for the Joback and Reid method, 7.67K for Stein and Brown, 12.09K for Marrero and Gani, 10.74 for Marrero and Pardillo and 6.30K for Cordes and Rarey). Overall, the proposed method yielded an average absolute deviation of 6.50K (1.52%) for a set of 2820 components. For the available methods, Joback and Reid produced an average absolute deviation of 21.37K (4.67%) for a set of 2514 components, 14.46K (3.53%) for 2578 components for Stein and Brown, 13.22K (3.15%) for 2267 components for Constantinou and Gani, 10.23 (2.33%) for 1675 components for Marrero and Pardillo and 8.18K (1.90%) for 2766 components for Cordes and Rarey. This implies that the proposed method yielded the lowest average deviation with the broadest range of applicability.

Chapter Eight

Recommendations

This work involved the development of a group contribution method for the estimation of the normal boiling point. In the development, the first step was to analyze the available methods to suggest the best approach for group contribution method. The next step involved finding a research strategy or procedure for the development of the group contribution method. To do this, however, the key aspects were the software tools and utilities described in Chapter 4. This meant that a significant amount of time was spent on these features. However, the major advantage of these features is that it can be readily applied in the development of any other property.

Implementing the above approach and research strategy, a group contribution method can be developed for the following properties:

- Critical properties
- Vapour pressure
- Normal melting point
- Standard Gibbs energy of formation at 298K
- Standard enthalpy of formation at 298K
- Standard enthalpy of vaporisation at 298K
- Standard enthalpy of vaporisation at the normal boiling point
- Standard enthalpy of vaporisation
- Standard entropy of vaporisation
- Standard entropy of vaporisation at 298K
- Standard enthalpy of fusion
- Heat capacity of an ideal gas
- Heat capacity of liquids
- Heat capacity of liquids at 298K
- Liquid viscosity

- Gas viscosity
- Acentric factor
- Liquid density
- Liquid volume at the normal boiling point
- Liquid molar volume at 298K
- Thermal conductivity
- Upper flammability limit
- Surface tension
- Water solubility
- Second virial coefficient
- 1-Octanol/water partition coefficient

Realising the many automotive and developmental procedures in this work, the time-span for the development of the above properties will generally depend on the available experimental data set. For example, the data set for critical properties is relatively small (less than 800 components for each critical property). Consequently, the time-span for a group contribution method for these properties will be relatively small.

A large number of the available group contribution methods for the above properties require critical properties, for example, vapour pressure. Considering the availability of critical properties, these methods are quite restricted. In all probability, the authors of these methods developed a group contribution method with a relatively high average deviation, and thus, included critical properties for a better estimation. Considering this, the best approach would be the inclusion of molecular properties, which is more easily accessible from a molecular simulation package, rather than critical properties. The best approach to develop this type of estimation method would be the procedure suggested in this work. This involves developing a group contribution method, whereby its limitations can be captured by the inclusion of molecular properties.

If the inclusion of molecular properties in a group contribution estimation method proves successful, then an investigation can be carried out to develop this type of estimation method for the prediction of ionic liquids or electrolyte solutions. Ionic liquids are a fairly recent development which serves as solvents in reaction chemistry, which has been around for the past 15 years. The major characteristic of ionic liquids in thermodynamics is that it does not have a measurable vapour pressure and thus can serve as an ideal solvent. However, the availability of experimental data is quite limited and considering the complexity of these electrolyte solutions, a group contribution estimation method would be quite restricted. This is the general reason that the inclusion of molecular properties would be more appropriate to strengthen the predictive capability. Typical ionic liquids properties are:

- Normal melting point
- Liquid and solid density
- Liquid viscosity
- Surface tension.

References

Agarwal R., Li Y., Santollani O., Satyro M. A., "Uncovering the Realities of Simulation – Part I," *CEP,* 42-52 May (2001a).

Agarwal R., Li Y., Santollani O., Satyro M. A., "Uncovering the Realities of Simulation – Part II," *CEP,* 64-72 June (2001b).

Atkins P. W., *Physical Chemistry*, 5th Edition, Oxford University Press, Oxford (1994).

Barrow G. M., *Physical Chemistry*, 4th ed., McGraw-Hill, Singapore 1985

Basarova P., Svoboda V., "Prediction of the enthalpy of vaporization by the group contribution method," *Fluid Phase Equilibria,* 105, 27-47 (1995).

Boyd D. B., *Ullman's Encyclopedia of Industrial Chemistry - Molecular Modelling,* 6th Edition, Wiley-VCH Verlag GmbH, Weinheim (2001).

Boethling R. S., Mackay D., *Handbook of Property Estimation Methods for Chemicals*, CRC Press, Boca Raton (2000).

Bruice P. Y., *Organic Chemistry,* 2nd Edition, Prentice Hall, New Jersey (1998)

Cavett R. H., "Physical data for Distillation Calculations: Vapor-Liquid Equilibria," *API Proc.,* Sec. III, 42 (1962).

Connolly M. L., http://www.netsci.org/Science/Compchem/feature14.html, (1996)

Constantinou L., "Estimation of Properties of Acyclic Organic Compounds through Conjugation," PhD Diss., Univ. of Maryland, College Park (1993).

Constantinou L., Gani R., “A New Group Contribution Method for the Estimation of Properties of Pure Hydrocarbons,” IVC-SEP 9319, Institut for Kemiteknik, The Technical Univ. of Denmark (1993).

Constantinou L., Gani R., “New Group Contribution Method for Estimating Properties of Pure Compounds,” *AIChE J*. 40 (10), 1697–1710 (1994a).

Constantinou L., Prickett S.E., Mavrovouniotis M.L., “Estimation of Properties of Acyclic Organic Compounds using Conjugation Operators,” *Ind. Eng. Chem. Res*., 39, 395 (1994b).

Constantinou L., Prickett S.E., Mavrovouniotis M.L., “Estimation of Thermodynamical and Physical Properties of Acyclic Hydrocarbons Using the ABC Approach and Conjugation Operators,” *Ind. Eng. Chem. Res*., 32(8), 1734 (1993).

Cordes W., Rarey J., “A New Method for the Estimation of the Normal Boiling Point of Non-Electrolyte Organic Compounds,” *Fluid Phase Equilibria* 201, 409-433 (2002).

Cordes W., Rarey J., Delique F., Gmehling J., “Software Development in Chemistry 7, Proceedings of Computer in der Chemie 7,” D. Ziessow (Ed.), Springer-Verlag, Berlin (1993).

Coutsikos P., Voutsas E., Magoulas K., Tassios D. P., “Prediction of vapour pressure of solid organic compounds with a group-contribution method,” *Fluid Phase Equilibria,* 207, 263-281 (2003).

Derr E.L., Deal C.H. Jr., “Analytical Solution of Groups,” *Inst. Chem. Eng. Symp. Ser.* (London), 3, 40 (1969)

Design Institute for Physical Property Data (AIChE Design Institute for Physical Properties), New York (1969)

Dortmund Data Bank, DDBST GmbH, Oldenburg (1973)

DDBST GmbH, *Artist 98 Manual,* Oldenburg (2000)

DDBST GmbH, *PureComponentProperties Manual,* Oldenburg (2001)

Elliot J. R., Lira C. T., *Introduction to Chemical Engineering Thermodynamics,* Prentice Hall, New Jersey (1999).

Ericksen D., Wilding W. V., Oscarson J. L., Rowley R. L., "Use of the DIPPR Database for Development of QSPR Correlations: Normal Boiling point," J. Chem. Eng. Data, 47 (2002) 1293-1302

Forsythe G.E., Malcolm M.A., Moler C.B., *Computer Methods for Mathematical Computations*, Prentice Hall, Englewood Cliffs, NJ (1977)

Fredenslund A., Gmehling J., Rasmussen P., *Vapor Liquid Equilibria using INIFAC*, Elsevier Scientific, Amsterdam (1977).

Hart H., Craine L. E., Hart D. J., *Organic Chemistry – A Short Course,* 9th Edition, Houghton Mifflin, Boston (1995).

Horvath A. L., *Molecular Design: Chemical Structure Generation from the Properties of Pure Organic Compounds,* Elsevier, (1992).

Joback K.G., Reid R.C., "Estimation Of Pure-Component Properties From Group-Contributions," *Chem. Eng. Commun.* 57 (1987) 233–243.

Katritzky A. R., Mu L., Lobanov V. S., "Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Organics," *J. Phys. Chem.,* 100, 10400-10407 (1996).

Klincewicz K. M., Reid R. C., "Estimation of Critical Properties with Group Contribution Methods," *AIChE Journal,* 30 (1), 137-142 (1984).

Krzyzaniak J. F., Myrdal P. B., Simamora P., Yalkowsky S. H., "Boiling Point and Melting Point Prediction for Aliphatic, Non-Hydrogen-Bonding Compounds," *Ind. Eng. Chem. Res.,* 34, 2530-2535 (1995).

Lydersen, A.L., Estimation of Critical Properties of Organic Compounds, Univ. Coll. Exp. Stn., Rept., Madison, WI, April, 1955

Lyman A. L., Reehl W. F., Rosenblatt D. H., *Handbook of Chemical Property Estimation Methods,* American Chemical Society, Washington, DC (1990).

Marrero-Morejon J., Gani R., "Group-contribution based estimation of pure components properties," *Fluid Phase Equilibria* 183-184 (2001) 183-208

Marrero-Morejon J., Pardillo-Fontdevila E., "Estimation of Pure Compound Properties Using Group-Interaction Contributions," *AIChE J.* 45 (3), 615-621 (1999).

Moura C. A. D., Carneiro H. P., "Common Difficluties in the Use of Process Simulators," *B. Tech. Petrobas,* 34 (3/4) July/December (1991).

Nelder J. A., Mead R., "A simplex method for function minimization," *Computer J.*, 308-313 (1965)

Pardillo-Fontdevila E., Gonzalez-Rubio R., "A Group-Interaction Contribution Approach. A New Strategy for the Estimation of Physical-Chemical Properties of Branched Isomers," *Chem. Eng. Commun.*, 163, 245 (1997)

Poling B. E., Prausnitz J. M., O'Connell J. P., *The Properties of Gases and Liquids,* 5th ed., McGraw-Hill, New York (2001).

Prausnitz J. M., Lichtenthaler R. N., de Azevedo E. G., *Molecular Thermodynamics of Fluid Phase Equilibria,* 3rd Edition, Prentice-Hall, Upper Saddle River, (1999).

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., *Numerical Recipes in C – The Art of Scientific Computing,* 2nd Edition, Cambridge University Press (2002).

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., *Numerical Recipes in Fortran 77 – The Art of Scientific Computing,* 2nd Edition, Cambridge University Press (1997).

Rarey J., *Computer Application in Technical Chemistry,* Wintersemester, Oldenburg (2001).

Rarey J., *VBA-Programming and DDB/DDBSP,* Oldenburg (2001)

Reid R. C., Prausnitz J. M., Poling B. E., *The Properties of Gases and Liquids,* 4th ed., McGraw-Hill, New York (1987).

Reid R. C., Sherwood T. K., *The Properties of Gases and Liquids – Their estimation and Correlation,* McGraw-Hill, New York (1958).

Retzekas E., Voutsas E., Magoulas K., Tassios D., "Prediction of Physical Properties of Hydrocarbons, Petroleum and Coal Liquid Fractions," *Ind. Eng. Chem. Res.*, 41, 1695-1702 (2002).

Sastri S. R. S., Rao K. K., "A new group contribution method for predicting viscosity of organic liquids," *The Chemical Engineering Journal,* 50, 9-25 (1992).

Simamora P., Yalkowsky S. H., "Group Contribution Methods for Predicting the Melting Points and Boiling Points of Aromatic Compounds," *Ind. Eng. Chem. Res.* 33, 1405-1409 (1994).

Stein S. E., Brown R. L., "Estimation of Normal Boiling Points from Group Contribution,“ *J. Chem. Inf. Comput. Sci.* 34 (1994) 581–587.

Tu C., Liu C., "Group-contribution estimation of the enthalpy of vaporization of organic compounds," *Fluid Phase Equilibria,* 121, 45-65 (1996).

Van Der Puy M., "A boiling point estimation method for alkanes and perfluoroalkanes," *Journal of Fluorine Chemistry,* 63, 165-172 (1993).

Van Ness H. C., Abbott M. M., *Classical Thermodynamics of Non Electrolyte Solution with Applications to Phase Equilibria*, McGraw Hill, New York (1982)

Voutsas E., Lampadariou M., Magoulas K., Tassios D., "Prediction of vapour pressures of pure compounds from knowledge of the normal boiling point," *Fluid Phase Equilibria,* 198, 81-93 (2002).

Vetere A., "Methods to predict the vaporization enthalpies at the normal boiling point temperature of pure compounds revisited," *Fluid Phase Equilibria,* 106, 1-10 (1995).

Wen X., Qiang Y., "Group Vector Space (GVS) Method for Estimating Boiling and Melting Points of Hydrocarbons," *J. Chem. Eng. Data,* 47, 286-288 (2002a).

Wen X., Qiang Y., "Group Vector Space Method for Estimating Melting and Boiling Points of Organic Compounds," *Ind. Eng. Chem. Res.*, 41, 5534-5537 (2002b).

Wessel M. D., Jurs P. C., "Prediction of Normal Boiling Points for a Diverse Set of Industrially Important Organic Compounds from Molecular Structure," *J. Chem. Inf. Comput. Sci.,* 35, 841-850 (1995).

Whitten K. W., Davis R. E., Peck M. L., *General Chemistry with Qualitative Analysis,* 5th Edition, Saunders College Publishing, New York (1996).

Appendix A

Previous Group Contributions

Table A-1: Group Contributions for Joback and Reid (1987)

Group Name	Value (K)	Group Name	Value (K)
Non-ring increments			
$-CH_3$	23.58	$-CH_2-$	22.88
>CH-	21.74	>C<	18.25
$=CH_2$	18.18	=CH-	24.96
=C<	24.14	=C=	26.15
≡CH	9.2	≡C-	27.38
Ring increments			
$-CH_2-$	27.15	>CH-	21.78
>C<	21.32	=CH-	26.73
=C<	31.01		
Halogen increments			
-F	-0.03	-Cl	38.13
-Br	66.86	-I	93.84
Oxygen increments			
-OH	92.88	-OH (a)	76.34
-O- (c)	22.42	-O- (r)	31.22
>C=O (c)	76.75	>C=O (r)	94.97
O=CH-	72.24	-COOH	169.09
-COO-	81.10	=O (except as above)	-10.50
Nitrogen increments			
$-NH_2$	73.23	>NH (c)	50.17
>NH (r)	52.82	>N- (c)	11.74
-N= (c)	74.60	-N= (r)	57.55
=NH	83.08	-CN	125.66

$-NO_2$	152.54		
Sulphur increments			
-SH	63.56	-S- (c)	68.78
-S- (r)	52.10		

(a – aromatic, r – ring, c – chain)

Table A-2: First order Group Contributions for Constantinou and Gani (1994)

Group Name	Value (K)	Group Name	Value (K)
CH_3	0.8894	CH	0.9225
CH	0.6033	C	0.2878
CH_2=CH	1.7827	CH=CH	1.8433
CH_2=C	1.7117	CH=C	1.7957
C=C	1.8881	CH_2=C=CH	3.1243
ACH	0.9297	AC	1.6254
$ACCH_3$	1.9669	$ACCH_2$	1.9478
ACCH	1.7444	OH	3.2152
ACOH	4.4014	CH_3CO	3.5668
CH_2CO	3.8967	CHO	2.8526
CH_3COO	3.636	CH_2COO	3.3953
HCOO	3.1459	CH_3O	2.2536
CH_2O	1.6249	CH-O	1.1557
FCH_2O	2.5892	CH_3NH_2	3.1656
$CHNH_2$	2.5983	CH_3NH	3.1376
CH_2NH	2.6127	CHNH	1.578
CH_3N	2.1647	CH_2N	1.2171
$ACNH_2$	5.4736	C_5H_4N	6.28
C_5H_3N	5.9234	CH_2CN	5.0525
COOH	5.8337	CH_2Cl	2.9637
CHCl	2.6948	CCl	2.2073
$CHCl_2$	3.93	CCl_2	3.56
CCl_3	4.5797	ACCL	2.6293
CH_2NO_2	5.7619	$CHNO_2$	5.0767

$ACNO_2$	6.0837	CH_2SH	3.2914
I	3.665	Br	2.6495
CH≡C	2.3678	C≡C	2.5645
Cl-(C=C)	1.7824	ACF	0.9442
$HCON(CH_2)_2$	7.2644	CF_3	1.288
CF_2	0.6115	CF	1.1739
COO	2.6446	CCl_2F	2.8881
HCClF	2.3086	$CClF_2$	1.9163
F (except as above)*	1.0081	$CONH_2$	10.3428
$CON(CH_3)_2$	7.6904	$CON(CH_2)_2$	6.7822
$C_2H_5O_2$	5.5566	$C_2H_4O_2$	5.4248
CH_3S	3.6796	CH_2S	3.6763
CHS	2.6812	C_4H_3S	5.7093
C_4H_2S	5.826		

* The method is not applied to highly partial fluorinated compounds

(A – Aromatic)

Table A-3: Second order Group Contributions for Constantinou and Gani (1994)

Group Name	**Value (K)**	**Group Name**	**Value (K)**
$(CH_3)_2CH$	-0.1157	$(CH_3)_3C$	-0.0489
$CH(CH_3)CH(CH_3)$	0.1798	$CH(CH_3)C(CH_3)_2$	0.3189
$C(CH_3)_2C(CH_3)$	0.7273	3 membered ring	0.4745
4 membered ring	0.3563	5 membered ring	0.1919
6 membered ring	0.1957	7 membered ring	0.3489
$CH_n=CH_m-CH_p=CH_k$, k,n,m,p e (0,2)	0.1589	$CH_3-CH_m=CH_n$, m,n e (0,2)	0.0668
$CH_2CH_m=CH_n$, m,n e (0,2)	-0.1406	$CH-CH_m=CH_n$ or $C-CH_m=CH_n$, m,n e (0,2)	-0.09
Alicyclic side chain $C_{cyclic}C_m$, m >1	0.0511	CH_3CH_3	0.6884
CHCHO or CCHO	-0.1074	CH_3COCH_2	0.0224
CH_3COCH or CH_3COC	0.092	$C_{cyclic}(=O)$	0.558
ACCHO	0.0735	CHCOOH or CCOOH	-0.1552

ACCOOH	0.7801	CH_3COOCH or CH_3COOC	-0.2383
$COCH_2COO$ or COCHCOO or COCCOO	0.4456	CO-O-CO	-0.1977
ACCOO	0.0835	CHOH	0.5385
COH	0.6331	$CH_m(OH)CH_n(OH)$, m,n e (0,2)	1.4108
$CH_{mcyclic}$-OH, m e (0,1)	-0.069	$CH_m(OH)CH_n(NH_p)$, m, n, p e (0,3)	1.0682
$CH_m(NH_2)CH_n(NH_2)$, m,n e (0,2)	0.4247	$CH_{mcyclic}$-NH_p-$CH_{ncyclic}$, m,n,p e (0,2)	0.2499
CH_m-O-CH_n=CH_p, m,n,p e (0,2)	0.1134	AC-O-CH_m, m e (0,3)	-0.2596
$CH_{mcyclic}$-S-$CH_{ncyclic}$, m,n e (0,2)	0.4408	CH_m=CH_n-F, m,n e (0,2)	-0.1168
CH_m=CH_n-Br, m,n e (0,2)	-0.3201	CH_m=CH_n-I, m,n e (0,2)	-0.4453
ACBr	-0.6776	ACI	-0.3678

* Corrections for stress strained effects are treated similar to second-order groups.

Table A-4: Group (Bond) Contributions for Marrero and Pardillo (1999)

Group	T_b	T_b^*	Group	T_b	T_b^*
Interactions with CH_3- (via single bond)					
CH_3-	-20.82	61.28	-OH (a)	133.04	736.93
CH_2-	33.19	194.25	-O-	31.94	228.01
CH-	26.94	194.27	>CO	64.46	445.61
>C<	22.71	186.41	-CHO	89.34	636.49
=CH-	18.17	137.18	-COOH	186.44	1228.84
=C<	23.91	182.2	-COO (o)	58.87	456.92
>C< (r)	23.04	194.40	-COO (c)	65.95	510.65
>CH- (r)	25.68	176.16	NH_2-	62.14	443.76
>C- (r)	20.25	180.60	>NH	41.60	293.86
=C< (r)	19.61	145.56	>N-	23.78	207.75
-F	-9.96	160.83	-CN	150.14	891.15
-Cl	44.44	453.70	$-NO_2$	169.64	1148.58
-Br	71.94	758.44	-SH	74.44	588.31

-I	111.04	1181.44	-S-	53.24	409.85
Interactions with non-ring –CH_2- (via single bond)					
-CH_2-	22	244.88	-OH (a)	108.85	673.24
>CH-	19.78	244.14	-O-	25.03	243.37
>C<	22.17	273.26	>CO	50.77	451.27
=CH-	25.30	201.80	-CHO	88.49	648.70
=C<	27.34	242.47	-COOH	156.34	1180.39
≡C-	26.20	207.49	-COO (o)	49.04	475.65
>CH- (r)	14.60	238.81	-COO (c)	53.64	541.29
>C< (r)	19.06	260.00	NH_2-	70.84	452.30
=C< (r)	4.1	167.85	>NH	35.62	314.71
-F	6.27	166.59	>N-	18.11	240.08
-Cl	62.72	517.62	-CN	130.85	869.18
-Br	84.49	875.85	-NO_2	70.35	612.31
-I	107.75	1262.80	-SH	47.45	451.03
Interactions with non-ring >CH- (via single bond)					
>CH-	21.94	291.41	-OH (a)	84.70	585.99
>C<	31.03	344.06	-O-	14.40	215.94
=CH-	14.44	179.96	>CO	45.66	434.45
=C<	33.24	249.1	-CHO	78.46	630.07
>CH- (r)	21.15	295.33	-COOH	170.37	1270.16
=C- (r)	-5.51	132.66	-COO (o)	44.23	497.23
-F	-2.06	68.80	-NH_2	47.06	388.44
-Cl	47.08	438.47	>NH-	22.34	260.32
Interactions with non-ring >C< (via single bond)					
>C<	46.38	411.56	-Cl	33.83	360.79
=CH-	23.36	286.30	-Br	50.42	610.26
=C<	41.2	286.42	-OH (a)	76.39	540.38
>CH- (r)	25.89	340.00	-O-	23.46	267.26
=C- (r)	2.61	188.99	>CO	38.63	373.71
-F	-7.70	-16.64	-COOH	164.43	1336.54

Interactions with non-ring =CH$_2$ (via double bond)					
=CH$_2$	-35.36	51.13	=C<	13.11	215.27
=CH-	28.66	205.73	=C=	17.02	183.55
Interactions with non-ring =CH- (via double bond)					
=CH-	35.33	334.64	-Cl	42.28	370.60
=C<	36.03	354.41	-O-	14.95	204.81
=C=	40.99	316.46	-CHO	92.68	658.53
=CH-	20.63	174.18	-COOH	180.68	1245.86
=C<	36.87	228.38	-COO (o)	44.27	423.86
≡C-	22.58	174.39	-COO (c)	59.38	525.35
=C< (r)	18.08	184.2	-CN	117.18	761.36
-F	-32.32	5.57			
Interactions with non-ring =C< (via double bond)					
=C<	45.7	399.58	=C=	44.51	321.02
Interactions with non-ring =C< (via simple bond)					
=C<	29.92	220.88	-Cl	36.54	367.05
-F	-13.78	-37.99			
Interactions with non-ring =C= (via triple bond)					
=O	10.32	160.42			
Interactions with non-ring ≡CH (via triple bond)					
≡CH	-16.26	120.85	≡C-	22.20	222.40
Interactions with non-ring ≡C- (via triple bond)					
≡C-	49.36	333.26			
Interactions with ring –CH$_2$- (via single bond)					
-CH$_2$- (r)	25.62	201.89	-O- (r)	29.60	225.52
>CH- (r)	21.77	209.40	>CO (r)	61.01	451.74
>C< (r)	20.34	182.74	>NH (r)	39.47	283.55
=CH- (r)	31.27	218.07	-S- (r)	56.34	424.13
=C< (r)	9.91	106.21			
Interactions with ring >CH- (via single bond)			**Interactions with ring >C< (via single bond)**		
>CH- (r)	19.23	210.66	>C< (r)	20.52	348.23

>C< (r)	22.71	220.24	=C< (r)	-3.11	-25.81
-O- (r)	22.17	169.17	>C< (r)	33.67	550.72
>CH- (r)	20.17	242.01	=C< (r)	16.16	408.64
-OH (p)	81.38	597.82	-F	-0.85	41.35
Interactions with ring =CH- (via double bond)					
>CH- (r)	8.41	112.00	=N- (r)	311.44	221.55
=C< (r)	36.01	291.15			
Interactions with ring =CH- (via single bond)					
>CH- (r)	44.57	285.07	>NH (r)	68.48	420.54
=C< (r)	24.95	237.22	=N- (r)	49.83	321.44
-O- (r)	22.15	171.59	-S- (r)	45.58	348.00
Interactions with ring =C< (via double bond)					
=C< (r)	66.09	477.77	=N- (r)	43.35	334.09
Interactions with ring =C< (via single bond)					
=C< (r)	1.99	180.07	-O-	3.66	199.70
-O- (r)	14.56	134.23	>CO	38.88	437.51
=N- (r)	16.03	174.31	-CHO	92.60	700.06
=C< (r)	-32.07	153.05	-COOH	151.44	1232.55
-F	-8.96	-48.79	-COO (c)	23.85	437.78
-Cl	30.76	347.33	NH_2-	77.47	517.75
-Br	51.77	716.23	>NH	40.53	411.29
-I	90.04	1294.98	>N-	48.18	422.51
-OH (p)	64.74	456.25	-CN	92.74	682.19
Interaction with -Cl (via single bond)					
CO	54.79	532.24			
Interactions with –O- (via single bond)					
CO	42.16	367.83	=N- (r)	57.78	382.25
Interactions with non-ring >CO (via single bond)					
CO	83.64	734.86			
Interactions with -H (forming formaldehyde. Formic acid, …)					
-CHO	49.34	387.17	-COO-	44.47	298.12
-COOH	169.14	1022.45			

Interactions with –NH_2 (via single bond)			**Interactions with non-ring –S- (via single bond)**		
>NH	115.75	673.59	-S-	61.17	597.59

(a – non-aromatic, p – aromatic, c – interaction via Carbon, o – interaction via oxygen, r – ring, rr – interaction of a group in a different ring)

Table A-5: Group Contributions for Stein and Brown (1994)

Name	GC	Name	GC
Carbon increments			
-CH_3	21.98	-CH_2-	24.22
CH_2- (r)	26.44	>CH-	11.86
>CH- (r)	21.66	>C<	4.5
>C< (r)	11.12	=CH_2	16.44
=CH-	27.95	=CH- (r)	28.03
=C<	23.58	=C< (r)	28.19
aaCH	28.53	aaC-	30.76
aaaC	45.46	≡CH	21.71
≡C-	32.99		
Oxygen increments			
-OH	106.27	-OH (1)	88.46
-OH (2)	80.63	-OH (3)	69.32
-OH (a)	70.48	-O-	25.16
-O- (r)	32.98	-OOH	72.92
Carbonyl increments			
-CHO	83.38	>CO	71.53
>CO (r)	94.76	-C(O)O-	78.85
-C(r)(O)O(r)-	172.49	-C(O)OH	169.83
-C(O)ONH_2	230.89	-C(O)ONH-	225.09
-C(r)(O)ON(r)H-	246.13	-C(O)ON<	142.77
-C(r)(O)ON(r)<	180.22		
Nitrogen increments			
-NH_2	61.98	-NH_2 (a)	86.63
>NH-	45.28	>NH- (r)	65.5

>N-	25.78	>N- (r)	32.77
>NOH	104.87	>NNO	184.68
aN	39.88	=NH	73.4
=N-	31.32	=N- (r)	43.54
=N(r)N(r)H-	179.43	-N(r)=C(r)RN(r)H-	284.16
-N=NNH-	257.29	-N=N-	90.87
-NO	30.91	-NO_2	113.99
-CN	119.16	-CN (r)	95.43
Halogen increments			
-F	0.13	-F (r)	-7.81
-Cl	34.08	-Cl (primary)	62.63
-Cl (secondary)	49.41	-Cl (tertiary)	36.23
-Cl (a)	36.79	-Br	76.28
-Br (r)	61.85	-I	111.67
-I (r)	99.93		
Sulphur increments			
-SH	81.71	-SH (r)	77.49
-S-	69.42	-S- (r)	69
>SO	154.5	>SO_2	171.58
>CS	106.2	>CS (r)	179.26
Phosphorus increments			
-PH_2	59.11	>PH	40.54
>P-	43.75	>PO-	107.23
Silicon increments			
>SiH-	27.15	>Si<	8.21
>Si< (r)	-12.16		
Miscellaneous increments			
>B-	-27.27	-Sc-	92.06
>Sn<	62.89		

(a – aromatic bond, r – ring, c – chain, 1 – primary, 2 – secondary, 3 - tertiary)

Table A-6: First order Group Contributions for Marrero and Gani (2001)

Group Name	Value (K)	Group Name	Value (K)
CH_3	0.8491	CH_2	0.7141
CH	0.2925	C	-0.0671
CH_2=CH	1.5596	CH=CH	1.5597
CH_2=C	1.3621	CH=C	1.2971
C=C	1.2739	CH_2=C=CH	2.684
CH_2=C=C	2.4014	CH=C=CH	2.54
CH≡C	1.7618	C≡C	1.6767
aCH	0.8365	aC fused with aromatic ring	1.7324
aC fused with non-aromatic subring	1.1995	aC (except as above)	1.5468
aN in aromatic ring	1.3977	aC-CH_3	1.5653
aC-CH_2	1.4925	aC-CH	0.8665
aC-C	0.5229	aC-CH=CH_2	2.4308
aC-CH=CH	2.9262	aC-C=CH_2	2.1472
aC-C=CH	2.3057	aC-C≡C	2.7341
OH	2.567	aC-OH	3.3205
COOH	5.1108	aC-COOH	6.0677
CH_3CO	3.1178	CH_2CO	2.6761
CHCO	2.1748	CCO	1.7287
aC-CO	3.465	CHO	2.5388
aC-CHO	3.5172	CH_3COO	3.1228
CH_2COO	2.985	CHCOO	2.2869
CCOO	1.6918	HCOO	2.5972
aC-COO	3.1952	aC-OOCH	0.4621
aC-OOC	3.0854	COO (except as above)	2.1903
CH_3O	1.7703	CH_2O	2.4217
CH-O	0.8924	C-O	0.4983
aC-O	1.8522	CH_2NH_2	2.7987
$CHNH_2$	2.0948	CNH_2	1.6525
CH_3NH	2.2514	CH_2NH	1.875

CHNH	1.2317	CH$_3$N	1.3841
CH$_2$N	1.1222	aC-NH$_2$	3.8298
aC-NH	2.923	aC-N	2.1918
NH$_2$ (except as above)	2.0315	CH=N	1.5332
C=N	1.4291	CH$_2$CN	4.5871
CHCN	3.9774	CCN	2.887
aC-CN	4.1424	CN (except as above)	3.0972
CH$_2$NCO	3.4891	CHNCO	3.122
aC-NCO	3.1853	CH$_2$NO$_2$	4.5311
CHNO$_2$	3.8069	CNO$_2$	3.3059
aC-NO$_2$	4.575	NO$_2$ (except as above)	3.2069
ONO	1.8896	ONO$_2$	3.2656
HCON(CH$_2$)$_2$	5.8779	HCONHCH$_2$	7.4566
CONH$_2$	6.5652	CONHCH$_3$	5.0724
CONHCH$_2$	6.681	CON(CH$_3$)$_2$	6.007
CON(CH$_2$)$_2$	5.0664	CONHCO	7.6172
CONCO	5.6487	aC-CONH$_2$	8.3775
aC-NH(CO)H	7.3497	aC-N(CO)H	5.1373
aC-CONH	7.585	aC-NHCO	7.4955
NHCONH	8.9406	NH$_2$CONH	16.3539
NH$_2$CON	2.0796	NHCON	7.1529
NCON	4.1459	aC-NHCONH$_2$	5.7604
aC-NHCONH	1.1633	CH$_2$Cl	2.6364
CHCl	2.0246	CCl	1.7049
CHO$_2$	3.342	CCl$_2$	2.9609
CCl$_3$	3.9093	CH$_2$F	1.5022
CHF	1.3738	CF	1.0084
CHF$_2$	2.2238	CF$_2$	0.5142
CF$_3$	1.1916	CO$_2$F	2.5053
HCOF	2.0542	CaF$_2$	1.7227
aC-Cl	2.0669	aC-F	0.7945
aC-I	3.7739	aC-Br	2.8414

I- (except as above)	3.1778	Br- (except as above)	2.4231
F- (except as above)	0.8504	Cl- (except as above)	1.5147
CHNOH	4.5721	CNOH	4.0142
OCH_2CH_2OH	4.8721	$OCHCH_2OH$	4.2329
OCH_2CHOH	3.6653	O-OH	3.1669
CH_2SH	3.1974	CHSH	2.591
CSH	2.0902	aC-SH	3.2675
SH (except as above)	2.3323	CH_3S	2.9892
CH_2S	2.6524	CHS	2.0965
CS	1.6412	aC-S-	2.9731
SO	6.2796	SO_2	7.0976
SO_3 (sulfite)	3.9199	SO_3 (sulfonate)	6.7785
SO_4 (sulfate)	5.5627	aC-SO	6.1185
aC-SO_2	8.4333	PH (phosphine)	2.0536
P (phosphine)	1.0984	PO_3 (phosphite)	2.79
PHO_3 (phosphonate)	5.6433	PO_3 (phosphonate)	4.5468
PHO_4 (phosphate)	5.1567	PO_4 (phosphate)	3.7657
aC-PO_4	2.3522	aC-P	2.9272
CO_3 (carbonate)	2.8847	C_2H_3O	2.8451
C_2H_2O	2.6124	C_2O	2.2036
CH_2 (cyc)	0.8234	CH (cyc)	0.5946
C(cyc)	0.0386	CH=CH (cyc)	1.5985
CH=C (cyc)	1.2529	C=C (cyc)	1.1975
CH_2=C (cyc)	1.5109	NH (cyc)	2.1634
N (cyc)	1.6541	CH=N (cyc)	6.523
C=N (cyc)	6.671	O (cyc)	1.0245
CO (cyc)	2.8793	S (cyc)	2.3256

(a – aromatic, cyc – cyclic)

Table A-7: Second order Group Contributions for Marrero and Gani (2001)

Group Name	**Value (K)**	**Group Name**	**Value (K)**
$(CH_3)_2CH$	-0.0035	$(CH_3)_3C$	0.0072

CH(CH$_3$)CH(CH$_3$)	0.316	CH(CH$_3$)C(CH$_3$)$_2$	0.3976
C(CH$_3$)$_2$C(CH$_3$)$_2$	0.4487	CH$_n$=CH$_m$-CH$_p$=CH$_k$ (k, m, n, p in 0..2)	0.1097
CH$_3$-CH$_m$=CH$_n$ (m, n in 0..2)	0.0369	CH$_2$-CH$_m$ =CH$_n$ (m,n in 0..2)	-0.0537
CH$_p$-CH$_m$=CH$_n$ (m, n in 0..2; p in 0..1)	-0.0093	CHCHO or CCHO	-0.1286
CH$_3$COCH$_2$	-0.0215	CH$_3$COCHorCH$_3$COC	-0.0803
CHCOOH or CCOOH	-0.3203	CH$_3$COOCH or CH$_3$COOC	-0.2066
CO-O-CO	-0.05	CHOH	-0.2825
COH	-0.5325	CH$_3$COCH$_n$OH (n in 0..2)	-0.2987
NCCHOH or NCCOH	0.2981	OH-CH$_n$-COO (n in 0..2)	-0.231
CH$_m$(OH)CH$_n$(OH) (m, n in 0..2)	0.8854	CH$_m$(OH)CH$_n$(-) (m, n, in 0..2)	0.5082
CH$_m$(NH2)CH$_n$(NH2) (m, n in 0..2)	-0.0064	CH$_m$(NH)CH$_n$(NH$_2$) (m, n in 1..2)	0.2318
HOOC-CH$_n$-COOH (n in 1..2)	-0.1222	HOOC-CH$_n$-CH$_m$-COOH (n, m in 1..2)	0.7686
HO-CH$_n$-COOH (n in 1..2)	-0.4625	CH$_3$-O-CH$_n$-COOH (n in 1..2)	0.9198
HS-CH$_n$-CH$_m$-COOH (n, m in 1..2)	-0.2697	NC-CH$_n$-CH$_m$-CN (n, m in 1..2)	1.8957
OH-CH$_n$-CH$_m$-CN (n, m in 1..2)	1.3434	HS-CH$_n$-CH$_m$-SH (n, m in 1..2)	0.1815
COO-CH$_m$-CH$_n$-OOC (n, m in 1..2)	0.3401	OOC-CH$_m$-CH$_n$-COO (n, m in 1..2)	0.5794
NC-CH$_n$-COO (n in 1..2)	1.2171	COCH$_n$COO (n in 1..2)	0.2427
CH$_m$-O-CH$_n$=CH$_p$ (m, n, p in 0..3)	0.1399	CH$_m$=CH$_n$-F (m, n in 0..2)	0.0591
CH$_m$=CH$_n$-Br (m, n in 0..2)	-0.3192	CH$_m$=CH$_n$-I (m, n in 0..2)	-0.3486
CH$_m$=CH$_n$-O (m, n in 0..2)	-0.0268	CH$_m$=CH$_n$-CN (m, n in 0..2)	0.0653
CH$_n$=CH$_m$-COO-CH$_p$ (m, n in 0..2)	-0.043	CH$_m$=CH$_n$-CHO (m, n in 0..2)	0.1102
CH$_m$=CH$_n$-COOH (m, n in 0..2)	0.0667	aC-CH$_n$-Cl (m, n in 0..2)	0.4537
aC-CH$_n$-NH$_m$ (m, n in 0..2)	0.259	aC-CH$_n$-O-(n in1..2)	-0.0425
aC-CH$_m$-OH (m, in 0..2)	0.1005	aC-CH$_m$-CN (m, in 0..2)	1.0587
aC-CH$_m$-CHO (m, in 0..2)	-0.0177	aC-CH$_m$-SH(m in 1..2)	0.1702
aC-CH$_n$-COOH (n in 0..2)	0.1584	aC-CH$_n$-CO- (n in 0..2)	0.3094
aC-CH$_n$-S- (n in 0..2)	0.103	aC-CH$_n$-OOC-H (n in 0..2)	0.2238
aC-CH$_n$-NO$_2$ (n in 0..2)	0.539	aC-CH$_n$-CONH$_2$ (n in 0..2)	-0.2197
aC-CH$_n$-OOC (n in 0..2)	0.0886	aC-CH$_n$-COO (n in 0..2)	0.0352

aC-CH(CH$_3$)$_2$	0.0196	aC-C(CH$_3$)$_3$	0.0494
aC-CF$_3$	-1.5974	(CH$_n$=C)$_{cyc}$-CHO (n in 0..2)	0.4267
(CH$_n$=C)$_{cyc}$-COO-CH$_m$ (m, n in 0..2)	0.0879	(CH$_n$=C)$_{cyc}$-CO- (n in 0..2)	0.6115
(CH$_n$=C)$_{cyc}$-CH$_3$ (n in 0..2)	0.0173	(CH$_n$=C)$_{cyc}$-CH$_2$ (n in 0..2)	-0.0504
(CH$_n$=C)$_{cyc}$-CN (n in 0..2)	-0.2474	(CH$_n$=C)$_{cyc}$-Cl (n in 0..2)	-0.5736
CH$_{cyc}$-CH$_3$	-0.121	CH$_{cyc}$-CH$_2$	-0.0148
CH$_{cyc}$-CH	0.1395	CH$_{cyc}$-C	0.1829
CH$_{cyc}$-CH=CH$_m$ (m, in 0..2)	-0.1192	CH$_{cyc}$-C=CH$_n$ (n in 0..2)	-0.0455
CH$_{cyc}$-O	0.2667	CH$_{cyc}$-F	-0.1899
CH$_{cyc}$-OH	-0.3179	CH$_{cyc}$-NH$_2$	-0.3576
CH$_{cyc}$-NH-CH$_n$ (n in 0..2)	-0.7458	CH$_{cyc}$-SH	-0.0569
CH$_{cyc}$-CN	0.4649	CH$_{cyc}$-COOH	0.1506
CH$_{cyc}$-CO	0.13	CH$_{cyc}$-NO$_2$	0.654
CH$_{cyc}$-S-	0.0043	CH$_{cyc}$-CHO	-0.2692
CH$_{cyc}$-O-	-0.2787	CH$_{cyc}$-OOCH	-0.2107
CH$_{cyc}$-COO	0.0926	CH$_{cyc}$-OOC	-0.4495
C$_{cyc}$-CH$_3$	0.0722	C$_{cyc}$-CH$_2$	0.0319
C$_{cyc}$-OH	-0.6775	>N$_{cyc}$-CH$_3$	0.0604
>N$_{cyc}$-CH$_2$	-0.308	AROMRINGs1s2	-0.159
AROMRINGs1s3	0.0217	AROMRINGs1s4	0.1007
AROMRINGs1s2s3	-0.1647	AROMRINGs1s2s4	-0.1387
AROMRINGs1s3s5	-0.1314	AROMRlNGs1s2s3s4	0.2745
AROMRINGs1s2s3s5	0.1645	AROMRINGs1s2s4s5	0.0754
PYRJDlNEs2	-0.1196	PYRJDlNEs3	0.0494
PYRJDlNEs4	0.1344	PYRJDINEs2s3	0.0032
PYRJDlNEs2s4	-0.0817	PYRJDINEs2s5	-0.1564
PYRJDlNEs2s6	-0.5176	PYRJDlNEs3s4	0.5477
PYRJDlNEs3s5	0.3533	PYRJDINEs2s3s6	-0.3888

(a – aromatic, cyc – cyclic)

Table A-8: Third order Group Contributions for Marrero and Gani (2001)

Group Name	Value (K)	Group Name	Value (K)
HOOC-(CH$_n$)$_m$-COOH (m>2, n in 0..2)	1.6498	NH$_2$-(CH$_n$)$_m$-OH (m>2, n in 0..2)	1.075
OH-(CH$_n$)$_m$-OH (m>2, n in 0..2)	0.7193	OH-(CH$_p$)$_k$-O-(CH$_n$)$_m$-OH (m,k>2, n,p in 0..2)	1.1867
OH-(CH$_p$)$_k$-NH$_x$-(CH$_n$)$_m$-OH (m,k>2, n,p,x in 0..2)	0.2991	CH$_p$-O-(CH$_n$)$_m$-OH (m>2, n,p in 0..2)	-0.4605
NH$_2$-(CH$_n$)$_m$-NH$_2$ (m>2, n in 0..2)	0.006	NH$_k$-(CH$_n$)$_m$-NH$_2$ (m,k>2, n in 0..2)	-0.1819
SH-(CH$_n$)$_m$-SH (m>2, n in 0..2)	0.4516	NC-(CH$_n$)$_m$-CN(m>2) (m>2, n in 0..2)	1.344
aC-(CH$_n$=CH$_m$)$_{cyc}$ (m in 0..2)	-0.3741	aC-Cac (different rings)	-0.4961
aC-CH$_{ncyc}$ (n in 0..2)	-0.4574	aC-CH$_{mcyc}$ (m in 0..2)	-0.1736
aC-(CH$_n$)$_m$-aC (m>2, n in 0..2)	0.3138	aC-(CH$_n$)$_m$-CH$_{cyc}$ (m>2, n in 0..2)	0.5928
CH$_{cyc}$-CH$_{cyc}$ (different rings)	0.4387	CH$_{cyc}$-(CH$_n$)$_m$-CH$_{cyc}$ (m>2, n in 0..2)	0.5632
CH multi-ring	0.1415	aC-CH$_m$-aC (m in 0..2)	0.2391
aC-(CH$_n$=C$_m$)-aC (m,n in 0..2)	0.7192	aC-CO-aC (different rings)	1.0171
aC-CH$_m$-CO-aC (m in 0..2)	0.9674	aC-CO-(C=CH$_m$)$_{cyc}$ (m in 0..2)	0.1126
aC-CO-CO-aC (different rings)	0.9317	aC-CO$_{cyc}$ (fused rings)	0.5031
aC-S$_{cyc}$ (fused rings)	0.2242	aC-S-aC (different rings)	0.0185
aC-SO$_n$-aC (different rings) (n in 0..4)	-0.085	aC-NH$_{ncyc}$ (fused rings) (n in 0..1)	1.1457
aC-NH-aC (different rings)	0.5768	aC-(C=N}$_{cyc}$ (different rings)	-0.5335
aC-(N=CH$_n$)$_{cyc}$ (fused, n in 0..1)	-5.2736	aC-O-CH$_n$-aC (different, n in 0..2)	0.6571
aC-O-aC (different rings)	-0.8252	aC-CH$_n$-O-CH$_m$-aC (n,m in 0..2)	0.279
aC-O$_{cyc}$ (fused rings)	-0.6848	AROMFUSED[2]	0.0441
AROMFUSED[2]s1	-0.1666	AROMFUSED[2]s2	-2.692
AROMFUSED[2]s2s3	-0.2807	AROMFUSED[2]s1s4	-0.3294
AROMFUSED[2]s1s2	-2.931	AROMFUSED[2Js1s3	-0.336

AROMFUSED[3]	0.0402	AROMFUSED[4a]	1.0466
AROMFUSED[4a]s1	-7.8521	AROMFUSED[4p]	0.9126
PYRlDlNE.FUSED[2]	-0.9432	PYRlDlNE.FUSED[2-iso]	-0.5844
PYRlDlNE.FUSED[4]	0.1733		

(a – aromatic, cyc – cyclic)

Table A-9: Group Contributions for Cordes and Rarey (2002)

Group Name	Value (K)	Group Name	Value (K)
F-(C,Si)	129.511	F-(C-([F,Cl]))-x	111.411
F-(C-([F,Cl]))-y	65.9125	F-(C-([F,Cl]$_2$))	144.464
F-(C(a))	-21.348	-CF=C<	73.5088
Cl- (C,Si)	327.158	Cl-((C,Si)-([F,Cl]))	300.288
Cl-((C,Si)-([F,Cl]$_2$))	275.233	Cl-(C(a))	204.105
-CCl=C<	299.52	COCl-	837.687
Br-(C/Si(na))	427.56	Br-(C(a))	351.895
I-(C,Si)	564.102	-OH short chain	515.544
-OH tertiary	401.033	HO-((C,Si)H$_2$-(C,Si)-(C,Si)-)	477.583
-OH (Ca)	354.061	HO-((C,Si)$_2$H-(C,Si)-(C,Si)-)	411.08
(C,Si)-O-(C,Si)	158.793	(C(a))-O(a)- (C(a))	79.2981
CHO-(C)	626.216	O=C<(C)$_2$	654.008
O=C(-O-)$_2$	911.983	COOH -(C)	1124.49
HCOO -(C)	712.76	(C)-COO -(C)	697.228
-C(c)OO–	1230.21	>(OC$_2$)<	861.138
-CO-O-CO-	1431.22	(C)-S-S-(C)	874.273
SH-(C)	459.247	(C)-S-(C)	479.985
-S(a)-	309.872	(C)-SO$_2$-(C)	1502.35
SCN-(C)	1002.53	NH$_2$-(C,Si)	361.207
NH$_2$- (Ca)	468.458	(C,Si)-NH-(C,Si)	259.446
(C,Si)$_2$>N-(C,Si)	121.99	=N(a)- (R5)	430.782
=N(a)- (R6)	282.737	C≡N-(C)	804.356
-CONH$_2$	1479.27	-CONH-	1323.88
-CON<	1058.87	OCN-	671.441

ONC-	1082.68	O=N-O-(C)	532.35
NO_2-(C)	907.229	NO_2-(C(a))	775.752
NO_3-	964.373	PO(O-)$_3$	1267.28
$AsCl_2$-	1173.13	CH_3-(ne)	188.555
CH_3-(e)	282.015	CH_3-(a)	176.705
-C(c)H_2-	250.119	-C(r)H_2-	246.871
>C(c)H-	260.938	>C(r)H-	241.804
>C(c)<	273.544	>C(c)<(a)	210.93
>C(c)<(e)	278.135	>C(r)<	265.032
>C(r)<(Ca)	281.964	>C(r)<(e,c)	264.839
>C(r)<(e,r)	304.422	=C(a)H-	245.521
=C(a)<(ne)	322.825	(a)=C(a)<$_2$(a)	386.361
=C(a)<(e)	377.988	H_2C(c)=C<	437.399
>C(c)=C(c)<	516.817	>C(c)=C(c)<(C(a))	607.968
>C(r)=C(r)<	507.998	-(e)C(c)=C(c)<	521.597
HC≡C-	468.03	-C≡C-	556.785
>Si<	294.323	>Si<(e)	219.416
(C)$_2$>Ge<(C)$_2$	301.028	$GeCl_3$-	1280.52
(C)$_2$>Sn<(C)$_2$	525.228	B(O-)$_3$	594.43

(a – aromatic atom or neighbour, c – chain atom or neighbour, e – very electronegative neighbours (N, O, F, Cl), ne – not very electronegative neighbours (not N, O, F, Cl), r – ring atom or neighbour)

Table A-10: Group Corrections for Cordes and Rarey (2002)

Group Name	**Value (K)**	**Group Name**	**Value (K)**
Para Pair(s)	37.5096	Meta Pair(s)	3.5994
Ortho Pair(s)	-44.8024	5 Ring	-27.0458
3/4 Ring	-39.0849	One Hydrogen	-131.323
No Hydrogen	-167.799		

Appendix B

Group Definitions

Table B-1: Group definition for first-order groups.

Abbreviations:		
	(e)	- very electronegative neighbours (N, O, F, Cl)
	(ne)	- not very electronegative neighbours (not N, O, F, Cl)
	(na)	- non-aromatic atom or neighbour
	(a)	- aromatic atom or neighbour
	(c)	- atom or neighbour is part of a chain
	(r)	- atom or neighbour is part of a ring

Group	Description	Name	ID[a] PR[b]	Examples
Periodic Group 17				
Fluorine				
F-	F- connected to C or Si	F-(C,Si)	19 86	2-fluoropropane, trimethylfluorosilane
	F- connected to a C or Si already substituted with one F or Cl and one other atom	F-(C-([F,Cl]))-a	22 83	1-chloro-1,2,2,2-tetrafluoroethane[r124], difluoromethylsilane
	F- connected to C or Si already substituted with at least one F and two other atoms	F-(C-(F))-b	21 80	1,1,1-trifluoroethane 2,2,3,3-tetrafluoropropionic acid
	F- connected to C or Si already substituted with at least one Cl and two other atoms	F-(C-(Cl))-b	102 81	trichlorofluoromethane[r11], 1,1-dichloro-1-fluoroethane [r141b]

	F- connected to C or Si already substituted with two F or Cl	F-(C-([F,Cl]$_{2}$))	23 82	1',1',1'-trifluorotoluene, 2,2,2-trifluoroethanol, trifluoroacetic acid
	F- connected to an aromatic carbon	F-(C(a))	24 85	fluorobenzene, 4-fluoroaniline
	F- on a C=C (vinylfluoride)	-CF=C<	20 84	vinyl fluoride, trifluoroethene, perfluoropropylene
Chlorine				
Cl-	Cl- connected to C or Si not already substituted with F or Cl	Cl- (C,Si)	25 72	butyl chloride, 2-chloroethanol, chloroacetic acid
	Cl- connected to C or Si already substituted with one F or Cl	Cl-((C,Si)-([F,Cl]))	26 71	dichloromethane, dichloroacetic acid, dichlorosilane
	Cl- connected to C or Si already substituted with at least two F or Cl	Cl-((C,Si)-([F,Cl]$_{2}$))	27 69	ethyl trichloroacetate, trichloroacetonitrile
	Cl- connected to an aromatic C	Cl-(C(a))	28 73	chlorobenzene
	Cl- on a C=C (vinylchloride)	-CCl=C<	29 70	vinyl chloride
COCl-	COCl- connected to C (acid chloride)	COCl-	77 19	acetyl chloride, phenylacetic acid chloride
Bromine				
Br-	Br- connected to a non-aromatic C or Si	Br-(C/Si(na))	30 66	ethyl bromide, bromoacetone
Br-	Br- connected to an aromatic C	Br-(C(a))	31 67	bromobenzene
Iodine				
I-	I- connected to C or Si	I-(C,Si)	32 64	ethyl iodide 2-iodotoluene

Periodic Group 16				
Oxygen				
-OH	-OH for aliphatic chains with less than five C (cannot be connected to aromatic fragments)	-OH short chain $< C_5$	36 91	ethanol, propanediol
	-OH connected to C or Si substituted with one C or Si in an at least five C or Si containing chain (primary alcohols)	-OH $> C_4$	35 87	1-nonanol, tetrahydrofurfuryl alcohol, ethylene cyanohydrin
	-OH connected to a C or Si substituted with two C or Si in a at least three C or Si containing chain (secondary alcohols)	HO-((C,Si)$_2$H-(C,Si)-(C,Si)-)	34 89	2-butanol, cycloheptanol
	-OH connected to C which has 4 non hydrogen neighbours (tertiary alcohols)	-OH tertiary	33 90	tert-butanol, diacetone alcohol
	-OH connected to an aromatic C (phenols)	-OH (Ca)	37 88	phenol, methyl salicylate
-O-	-O- connected to 2 neighbours which are each either C or Si (ethers)	(C,Si)-O-(C,Si)	38 93	diethyl ether, 1,4-dioxane
	-O- in an aromatic ring with aromatic C as neighbours	(C(a))-O(a)-(C(a))	65 92	furan, furfural
-CHO	CHO- connected to non-aromatic C (aldehydes)	CHO-(Cna)	52 53	acetaldehyde, pentanedial
	CHO- connected to aromatic C (aldehydes)	CHO-(Ca)	90 52	furfural, benzaldehyde
>C=O	-CO- connected to two non-aromatic C (ketones)	O=C<(Cna)$_2$	51 55	acetone, methyl cyclopropyl ketone

	-CO- connected to two C with at least one aromatic C (ketones)	$(O=C<(C)_2)a$	92 54	acetophenone, benzophenone
	-CO connected to N	>N(C=O)-	109 39	methyl thioacetate
	-CO connected to two N (urea)	>N-(C=O)-N<	100 2	urea-1,1,3,3-tetramethyl
$O=C(-O-)_2$	Non-cyclic carbonate	$O=C(-O-)_2$	79 15	dimethyl carbonate
COOH -	-COOH connected to C	COOH -(C)	44 24	acetic acid
-COO -	HCOO- connected to C (formic acid ester)	HCOO -(C)	46 27	ethyl formate, phenyl formate
	-COO- connected to two C (ester)	(C)-COO -(C)	45 25	ethyl acetate, vinyl acetate
	-COO- in a ring, C is connected to C (lactone)	-C(c)OO–	47 26	ε-caprolactone, crotonolactone
-OCOO-	-CO connected to two O (Carbonates)	-OCOO-	103 34	propylene carbonate 1,3 dioxolan-2-one
-OCON<	-CO connected to O and N (carbamate)	-OCON<	99 1	trimethylsilyl methylcarbamate
$>(OC_2)<$	$>(OC_2)<$ (epoxide)	$>(OC_2)<$	39 50	propylene oxide
-CO-O-CO-	anhydride connected to two C	-C=O-O-C=O-	76 12	acetic anhydride, butyric anhydride
	cyclic anhydride connected to two C	(-C=O-O-C=O-)r	96 11	maleic anhydride, phthalic anhydride
-O-O-	Peroxide	-O-O-	94 32	di-tert-butylperoxide

Sulphur				
-S-S-	-S-S- (disulfide) connected to two C	(C)-S-S-(C)	55 51	dimethyldisulfide, 1,2-dicyclopentyl-1,2-disulfide
-SH	-SH connected to C (thioles)	SH-(C)	53 74	1-propanethiol
-S-	-S- connected to two C	(C)-S-(C)	54 75	methyl ethyl sulfide
	-S- in an aromatic ring	-S(a)-	56 76	thiazole, thiophene
-SO_2-	Non-cyclic sulfone connected to two C (sulfones)	(C)-SO_2-(C)	82 18	sulfolane, divinylsulfone
>SO_4	$S(=O)_2$ connected to two O (sulfates)	>SO_4	104 35	dimethyl sulfate
-SO_2N<	-$S(=O)_2$ connected to N	-SO_2N<	105 36	n,n-diethylmethanesulfonamide
>S=O	Sulfoxide	>S=O	107 37	1,4-thioxane-s-oxide tetramethylene sulfoxide
SCN-	SCN- (isothiocyanate) connected to C	S=C=N-(C)	81 20	allyl isothiocyanate
Selenium				
>Se<	>Se< connected to at least 1 C or Si	>Se<	116 46	dimethyl selenide
Periodic Group 15				
Nitrogen				
NH_2-	NH_2- connected to either C or Si	NH_2-(C,Si)	40 95	hexylamine, ethylenediamine
	NH_2- connected to an aromatic C	NH_2- (Ca)	41 94	aniline, benzidine

-NH-	-NH- connected to 2 neighbours which are each either C or Si (secondary amines)	(C,Si)-NH-(C,Si)	42 99	diethylamine, diallyl amine
	-NH- connected to 2 C or Si neighbours, with at least 1 ring neighbour (secondary amines)	(C,Si)r-NH-(Ca,Si)r	97 98	morpholine pyrrolidine
	-NH- connected to 2 C or Si neighbours, with at least 1 aromatic neighbour (secondary amines)	(C,Si)a-NH-(Ca,Si)a	98 99	diphenylamine n-methylaniline
>N<	>N- connected to 3 neighbours which are each either C or Si (tertiary amines)	(C,Si)$_2$>N-(C,Si)	43 100	n,n-dimethylaniline, nicotine
	>N- connected to 3 C or Si neighbours, with at least 1 aromatic neighbour (tertiary amines)	a(C,Si)$_2$>N-(C,Si)a	110 42	n,n-dimethylaniline n,n-diethylaniline
	Quaternary amine connected to 4 C or Si	(C,Si)$_2$>N<(C,Si)$_2$	101 33	n,n,n,n-tetramethylmethylenediamine
=N-	double bonded amine connected to at least 1 C or Si	(C,Si)=N-	91 101	acetonin
-N-	aromatic -N- in a 5 membered ring, free electron pair	=N(a)- (r5)	66 97	piperidine, thiazole
=N-	aromatic =N- in a 6 membered ring	=N(a)- (r6)	67 96	pyridine, nicotine
C≡N-	-C≡N (cyanide) connected to C	(C)-C≡N	57 56	acetonitrile, 2,2'-dicyano diethyl sulfide

	-C≡N (cyanide) connected to N	(N)-C≡N	111 41	dimethylcyanamide
	-C≡N (cyanide) connected to S	(S)-C≡N	108 38	methyl thiocyanate
CNCNC-r	imadizole	..=CNC=NC=..	106 3	1 methyl 1 imadizole
-CONH<	-CONH$_2$ (amide)	-CONH$_2$	50 28	acetamide
	-CONH- (monosubstituted amide)	-CONH-	49 48	n-methylformamide, 6-caprolactam
	-CON< (disubstituted amide)	-CON<	48 49	n,n-dimethylformamide (dmf)
OCN-	OCN- connected to C or Si (cyanate)	OCN-	80 29	butylisocyanate, hexamethylene diisocyanate
ONC-	ONC- (oxime)	ONC-	75 30	methyl ethyl ketoxime
-ON=	-ON= connected to C or Si (isoazole)	-ON=(C,Si)	115 45	isoazole 5-phenyl isoazole
NO$_2$-	nitrites (esters of nitrous acid)	O=N-O-(C)	74 23	ethyl nitrite, nitrous acid methyl ester
	NO$_2$- connected to aliphatic C	NO$_2$-(C)	68 21	1-nitropropane
	NO$_2$- connected to aromatic C	NO$_2$-(C(a))	69 22	nitrobenzene
NO$_3$-	nitrate (esters of nitric acid)	NO$_3$-	72 14	n-butylnitrate, 1,2-propanediol dinitrate

Phosphorous				
>P(O-)$_3$	phosphates with four O substituents	PO(O-)$_3$	73 10	triethyl phosphate, tri-(2,4-dimethylphenyl) phosphate
>P<	phosphorus connected to at least 1 C or S (phosphine)	>P<	113 43	triphenylphosphine trietylphosphane
Arsine				
AsCl$_2$-	AsCl$_2$ connected to C	AsCl$_2$-	84 17	ethylarsenic dichloride
Periodic Group 14				
Carbon				
-CH$_3$	CH3- not connected to either N, O, F or Cl	CH$_3$-(ne)	1 104	decane
	CH$_3$- connected to either N, O, F or Cl	CH$_3$-(e)	2 102	dimethoxymethane, methyl butyl ether
	CH$_3$- connected to an aromatic atom (not necessarily C)	CH$_3$-(a)	3 103	toluene, p-methyl-styrene
-CH$_2$-	-CH$_2$- in a chain	-C(c)H$_2$-	4 111	butane
	-CH$_2$- in a ring	-C(r)H$_2$-	9 112	cyclopentane
>CH-	>CH- in a chain	>C(c)H-	5 117	2-methylpentane
	>CH- in a ring	>C(r)H-	10 116	methylcyclohexane
>C<	>C< in a chain	>C(c)<	6 119	neopentane
	>C< in a chain connected to at least one aromatic carbon	>C(c)<(a)	8 108	ethylbenzene, diphenylmethane
	>C< in a chain connected to at least one F, Cl, N or O	>C(c)<(e)	7 107	ethanol

	>C< in a ring	>C(r)<	11 118	beta-pinene
	>C< in a ring connected to at least one aromatic carbon	>C(r)<(Ca)	14 106	indene, 2-methyl tetralin
	>C< in a ring connected to at least one N or O which are not part of the ring or one Cl or F	>C(r)<(e,c)	12 109	cyclopentanol, menthol
	>C< in a ring connected to at least one N or O which are part of the ring	>C(r)<(e,r)	13 110	morpholine, nicotine
=C(a)<	aromatic =CH-	=C(a)H-	15 105	benzene
	aromatic =C< not connected to either O,N,Cl or F	=C(a)<(ne)	16 115	ethylbenzene, benzaldehyde
	aromatic =C< with 3 aromatic neighbours	(a)=C(a)<$_{2}$(a)	18 114	naphthalene, quinoline
	aromatic =C< connected to either O,N,Cl or F	=C(a)<(e)	17 113	aniline, phenol
>C=C<	H2C=C< (1-ene)	H2C(c)=C<	61 58	1-hexene
	>C=C< (both C have at least one non-H neighbour)	>C(c)=C(c)<	58 63	2-heptene, mesityl oxide
	non-cyclic >C=C< connected to at least one aromatic C	>C(c)=C(c)< (C(a))	59 60	isosafrole, cinnamic alcohol
	cyclic >C=C<	>C(r)=C(r)<	62 61	cyclopentadiene
	non-cyclic >C=C< substituted with at least one F, Cl, N or O	-(e)C(c)=C(c)<	60 59	trans-1,2-dichloroethylene, perfluoroisoprene
-C≡C-	HC≡C- (1-ine)	HC≡C-	64 57	1-heptyne

	-C≡C-	-C≡C-	6362	2-octyne
>C=C=C<	cumulated double bond	>C=C=C<	876	1,2 butadiene dimethyl allene
>C=C-C=C<	conjugated double bond in a ring	>C=C-C=C<	887	cyclopentadiene abietic acid
>C=C-C=C<	conjugated double bond in a chain	>C=C-C=C<	898	isoprene 1,3 hexadiene
-C≡C-C≡C-	conjugated triple bond	-C≡C-C≡C-	959	2,4 hexadiyne
Silicon				
>Si<	>Si<	>Si<	7079	butylsilane
	>Si< connected to at least one O	>Si<(O)	7177	hexamethyl disiloxane
	>Si< connected to at least one F or Cl	>Si<(F,Cl)	7816	trichlorosilane,
Germanium				
>Ge<	>Ge< connected to four carbons	(C)$_2$>Ge<(C)$_2$	8668	tetramethylgermane
GeCl$_3$-	GeCl$_3$- connected to carbons	GeCl$_3$-	8513	fluorodimethylsilyl(trichlorogermanyl)methane
Stannium				
>Sn<	>Sn< connected to four carbons	(C)$_2$>Sn<(C)$_2$	8365	tetramethylstannane
Periodic Group 13				
Boron				
B(O-)$_3$	Non-cyclic boric acid ester	B(O-)$_3$	7816	triethyl borate

Aluminum				
>Al<	>Al< connected to at least 1 C or Si	>Al<	117 47	triethylaluminum

[a] ID – Identification Number, [b] PR – Priority Number

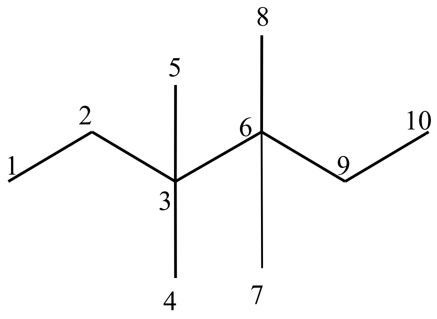
Table B-2: Group definition for second-order corrections.

Name	Description	ID	Examples
C=C-C=O	-C=O connected to sp^2 carbon	118	benzaldehyde furfural
(C=O)-C([F,Cl]$_{2,3}$)	Carbonyl connected to carbon with two or more halogens	119	dichloroacetyl chloride
(C=O)-(C([F,Cl]$_{2,3}$))$_2$	Carbonyl connected to two carbon with two or more halogens each	120	perfluoro-2-propanone
C-[F,Cl]$_3$	Carbon with three halogens	121	1,1,1-triflourotoluene
(C)$_2$-C-[F,Cl]$_2$	Secondary carbon with two halogens	122	2,2-dichloropropane
No Hydrogen	Component has no hydrogen	123	perfluoro compounds
One Hydrogen	Component has one hydrogen	124	nonafluorobutane
3/4 Ring	A three or four-membered non-aromatic ring	125	cyclobutene
5 Ring	A five-membered non-aromatic ring	126	cyclopentane
Ortho Pair(s)	Ortho position – Counted only once and only if there are no meta or para pairs	127	o-xylene
Meta Pair(s)	Meta position – Counted only once and only if there are no para or ortho pairs	128	m-xylene
Para Pair(s)	Para position – Counted only once and only if there are no meta or ortho pairs	129	p-xylene
((C=)(C)C-CC$_3$)	Carbon with four carbon neighbours and 1 double bonded carbon neighbour	130	Tert-butylbenzene
C$_2$C-CC$_2$	Carbon with four carbon neighbours, two on each side	131	bicyclohexyl
C$_3$C-CC$_2$	Carbon with five carbon neighbours	132	Ethyl bornyl ether
C$_3$C-CC$_3$	Carbon with six carbon neighbours	133	2,2,3,3-tetrametylbutane

Appendix C

Examples

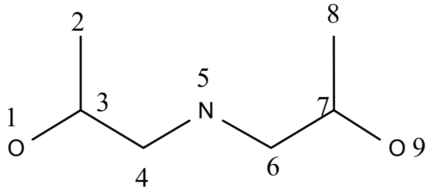
Table C-1: Estimation of the normal boiling temperature of 3,3,4,4-tetramethylhexane.

<table>
<tr><td colspan="2">Component: 3,3,4,4-
Tetramethylhexane
Number of atoms: 10</td><td colspan="3">.
 </td></tr>
<tr><td><b>Group</b></td><td><b>Atoms</b></td><td><b>Frequency</b></td><td><b>Contribution</b></td><td><b>Total</b></td></tr>
<tr><td>1</td><td>1,4,5,7,8,10</td><td>6</td><td>177.4949</td><td>1064.9694</td></tr>
<tr><td>4</td><td>2,9</td><td>2</td><td>239.7475</td><td>479.495</td></tr>
<tr><td>6</td><td>3,6</td><td>2</td><td>249.9094</td><td>499.8189</td></tr>
<tr><td colspan="5"><b>Steric Corrections</b></td></tr>
<tr><td><b>Group</b></td><td><b>Bond</b></td><td><b>Frequency</b></td><td><b>Contribution</b></td><td><b>Total</b></td></tr>
<tr><td>N6</td><td>3-6</td><td>1</td><td>121.3234</td><td>121.3234</td></tr>
<tr><td colspan="4"><b>Sum</b></td><td><b>2165.6067</b></td></tr>
</table>

$$T_b = \frac{2165.6067K}{10^{0.6587} + 1.6902} + 84.3359K = 431.0K$$

Experimental T_b = 444.0 K

Table C-2: Estimation of the normal boiling temperature of di-isopropanolamine.

Component: Di-Isopropanolamine Number of atoms: 9				
Group	**Atoms**	**Frequency**	**Contribution**	**Total**
1	2,3	2	177.4949	354.9898
7	3,4,6,7	4	267.1072	1068.4288
34	1,9	2	390.7067	781.4134
42	5	1	223.4973	223.4973
		Interactions		
Group	**Atoms**	**Frequency**	**Contribution**	**Total**
OH-OH	1-9	(1[a])/9	292.2832	32.4759
OH-NH	1-5, 9-5	(2[b])/9	287.5930	63.9095
		Sum		**2524.7149**

$$T_b = \frac{2524.7149\text{K}}{9^{0.6587} + 1.6902} + 84.3359\text{K} = 509.2\text{K}$$

Experimental T_b = 522.0 K

[a] I(OH-OH) = (OH_1 – OH_9) + (OH_9 – OH_1) = 2, $N_{OH\text{-}OH}$ = 2 / (3 - 1)* = 1

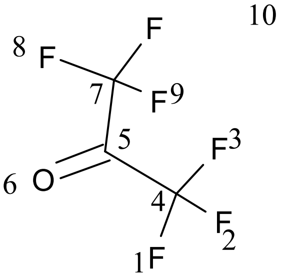
[b] I(NH-OH) = (OH_1 –NH_5) + (OH_9 – NH_5) + (NH_5 – OH_1) + (NH_5 – OH_9) = 4

$N_{NH\text{-}OH}$ = 4/(3 – 1)* = 2

* Total number of interaction groups minus the interaction with itself = (3 – 1).

NB: I – Interactions, N - Frequency

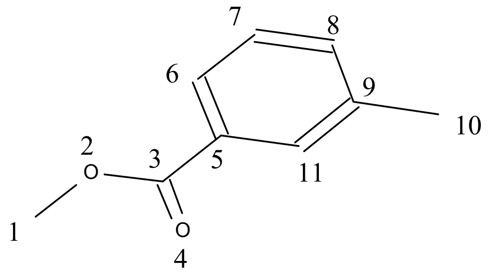
Table C-3: Estimation of the normal boiling temperature of perfluoro-2-propanone.

<table>
<tr><td colspan="2">Component: Perflouro-2-Propanone
Number of atoms: 10</td><td colspan="3">
</td></tr>
<tr><th>Group</th><th>Atoms</th><th>Frequency</th><th>Contribution</th><th>Total</th></tr>
<tr><td>7</td><td>4,7</td><td>2</td><td>267.1072</td><td>534.2144</td></tr>
<tr><td>21</td><td>1,2,3,8,9,10</td><td>6</td><td>53.2649</td><td>319.5894</td></tr>
<tr><td>51</td><td>5,6</td><td>1</td><td>619.5643</td><td>619.5643</td></tr>
<tr><th colspan="5">Corrections</th></tr>
<tr><th>Group</th><th>Atoms</th><th>Frequency</th><th>Contribution</th><th>Total</th></tr>
<tr><td>120</td><td>5</td><td>1</td><td>-248.0734</td><td>-248.0734</td></tr>
<tr><td>121</td><td>4,7</td><td>2</td><td>-20.3435</td><td>-40.687</td></tr>
<tr><td>123</td><td>-</td><td>1</td><td>-172.7072</td><td>-172.7072</td></tr>
<tr><th colspan="4">Sum</th><th>1011.9007</th></tr>
</table>

$$T_b = \frac{1011.9007\text{K}}{10^{0.6587} + 1.6902} + 84.3359\text{K} = 246.3\text{K}$$

Experimental T_b = 245.9 K

Table C-4: Estimation of the normal boiling temperature of methyl m-toluate.

Component: Methyl m-Toluate Number of atoms: 11		 7 8 6 9 10 2 O 3 5 11 1 O 4		
Group	**Atoms**	**Frequency**	**Contribution**	**Total**
2	1	1	251.7607	251.7607
3	10	1	158.0649	158.0649
15	6,7,8,11	4	235.7455	942.982
16	5,9	2	315.5639	631.1278
45	2,3,4	1	636.9917	636.9917
Corrections				
Group	**Atoms**	**Frequency**	**Contribution**	**Total**
128	5 & 9	1	-4.1349	-4.1349
118	3-5-6	1	40.8249	40.8249
Sum				**2657.6169**

$$T_b = \frac{2657.6169\text{K}}{11^{0.6587} + 1.6902} + 84.3359\text{K} = 490.5\text{K}$$

Experimental T_b = 492.3 K

Appendix D

Group Contributions

Table D-1: Group contribution values for first-order groups.

Group number	Group contribution (K)	Mean absolute error (%)	Mean absolute error (K)	Standard deviation (K)	Number of components
1	177.4949	1.35	5.85	7.87	1844
2	251.7607	1.63	6.96	9.35	272
3	158.0649	1.25	5.91	7.76	172
4	239.7475	1.19	5.39	7.12	1154
5	240.9746	1.15	5.29	7.04	391
6	249.9094	1.76	7.47	10.59	97
7	267.1072	1.65	6.99	9.26	1027
8	201.3488	1.06	5.56	7.04	183
9	239.7747	1.58	6.63	8.88	330
10	222.3732	1.28	5.51	7.51	156
11	210.2796	1.00	4.32	6.49	50
12	251.1733	1.36	5.64	8.05	78
13	291.4089	1.86	8.00	10.66	123
14	244.7234	1.38	7.26	9.90	25
15	235.7455	1.34	6.74	8.91	694
16	315.5639	1.26	6.44	8.57	543
17	349.0680	1.50	7.37	9.70	299
18	368.1391	1.09	6.12	7.83	64
19	106.7661	1.90	6.49	9.62	39
20	49.3515	3.80	8.84	10.07	9
21	53.2649	2.46	8.44	10.70	150

22	78.8553	2.48	8.27	10.54	29
23	103.6996	1.75	4.43	5.90	3
24	-19.9537	2.02	7.89	9.93	28
25	331.3170	2.00	8.22	10.70	120
26	287.5276	1.76	6.90	9.23	49
27	267.7401	2.21	8.60	11.53	55
28	205.5617	1.82	8.81	11.42	73
29	292.9088	1.98	6.90	9.29	35
30	419.9362	2.13	8.05	9.94	68
31	378.6803	1.31	6.32	8.44	26
32	557.0328	1.42	5.82	6.96	28
33	350.4125	1.45	6.30	8.48	49
34	390.7067	1.74	7.98	10.25	97
35	444.3332	1.44	7.16	9.35	89
36	488.6496	2.12	9.20	11.61	53
37	361.7430	2.07	10.37	12.98	63
38	146.7517	1.68	7.34	9.76	457
39	821.5838	1.15	4.61	6.47	20
40	321.5537	1.16	4.91	6.12	55
41	441.7401	1.20	6.26	7.73	38
42	223.4973	0.87	3.67	5.21	41
43	127.0204	2.12	9.49	12.12	51
44	1081.4627	1.67	8.57	11.30	60
45	636.9917	1.63	7.70	10.31	283
46	642.7711	1.07	4.56	8.23	19
47	1186.5677	0.80	3.85	4.50	3
48	1054.2086	1.87	8.87	11.89	10
49	1366.0383	1.50	7.53	10.21	6
50	1488.9557	1.50	7.43	8.16	4
51	619.5643	1.69	7.63	10.22	114

52	554.3118	1.79	7.24	10.44	36
53	434.4601	0.98	3.79	4.80	50
54	462.0433	1.35	5.58	7.12	56
55	865.4276	0.72	2.97	3.91	4
56	304.8520	1.47	6.83	9.76	31
57	720.0136	1.26	5.73	7.80	44
58	476.2776	1.53	6.11	8.03	115
59	586.2565	1.49	7.82	10.00	12
60	500.8132	2.05	7.10	8.86	29
61	413.0742	1.60	5.93	7.74	198
62	476.4552	2.05	8.47	10.73	91
63	512.7882	0.67	2.74	3.68	24
64	422.6805	1.80	6.24	8.12	28
65	37.0483	0.64	2.65	3.75	18
66	453.9585	1.52	6.81	8.25	23
67	307.2670	1.49	7.03	8.74	44
68	867.6357	4.26	16.38	18.47	5
69	822.1739	1.15	6.14	7.91	30
70	282.2966	1.39	5.01	7.11	37
71	208.1473	2.25	10.77	14.26	43
72	921.4868	0.86	3.48	4.03	6
73	1154.1140	0.89	4.97	5.12	4
74	494.8568	0.49	1.66	1.83	7
75	1042.1862	1.80	8.00	9.12	9
76	1252.6899	2.69	10.34	12.88	5
77	779.8043	1.33	5.94	8.18	28
78	541.2516	1.49	6.35	7.51	8
79	881.0107	0.37	1.70	1.96	4
80	661.1921	1.24	5.34	6.32	16
81	1019.7708	0.51	2.08	2.30	3

82	1561.7087	0.80	4.47	4.99	3
83	510.9847	0.29	1.14	1.34	3
84	1151.1951	0.72	3.17	3.68	6
85	1210.6067	0.30	1.20	1.44	3
86	348.1842	0.00	0.00	0.00	1
87	664.8292	3.31	9.20	11.63	5
88	958.8024	3.02	13.41	15.66	12
89	930.0789	2.80	10.12	12.54	22
90	560.5246	0.78	3.98	7.05	11
91	229.6735	1.42	5.90	7.01	6
92	606.3001	1.05	5.78	7.74	25
93	215.4991	2.30	8.39	11.33	37
94	273.7071	0.00	0.00	0.00	1
95	1219.5429	0.00	0.00	0.00	1
96	2083.9341	1.46	7.71	8.07	4
97	201.8072	1.57	6.16	8.21	19
98	381.2443	0.65	3.40	4.04	5
99	888.0460	1.30	6.13	7.50	11
100	1047.6115	0.00	0.00	0.00	1
101	-109.1249	1.37	5.34	5.98	3
102	111.2071	2.43	8.90	12.59	8
103	1575.4244	0.56	2.89	2.89	2
104	1485.2109	0.44	2.07	2.08	2
105	1508.8392	1.24	6.52	7.89	3
106	485.4450	1.93	9.34	10.34	4
107	1381.0220	1.33	7.49	7.49	2
108	660.4903	1.57	6.51	7.26	3
109	492.6546	0.94	3.91	4.73	4
110	194.1012	1.47	6.89	8.51	5
111	972.2145	0.00	0.00	0.00	1

113	429.0999	0.39	1.65	1.76	4
115	613.3196	1.17	5.05	5.14	3
116	562.8295	0.00	0.00	0.00	1
117	762.4185	1.30	5.50	5.54	2

Table D-2: Group contribution values for second-order corrections.

Group number	**Group contribution (K)**	**Mean absolute error (%)**	**Mean absolute error (K)**	**Standard deviation (K)**	**Number of components**
118	40.8249	1.20	6.06	8.22	135
119	-82.3645	2.96	12.39	13.89	19
120	-248.0734	0.16	0.43	0.43	2
121	-20.3435	2.36	8.33	10.70	139
122	15.6007	2.25	8.41	10.87	69
123	-172.7072	2.00	6.86	9.07	99
124	-99.9809	2.03	6.77	9.49	37
125	-62.4691	1.84	6.53	8.92	52
126	-40.0797	1.51	6.34	8.55	180
127	-29.5015	1.35	6.79	9.01	83
128	-4.1349	1.49	7.16	9.60	85
129	15.7780	1.37	6.77	8.79	102
130	25.7299	1.43	6.59	8.32	27
131	35.8705	1.20	5.37	7.08	88
132	51.9931	0.99	4.43	6.13	44
133	121.3234	0.72	3.23	4.42	17

Table D-3: Group contribution values for second-order group interactions.

Group number	**Group contribution (K)**	**Mean absolute error (%)**	**Mean absolute error (K)**	**Standard deviation (K)**	**Number of components**
OH - OH	292.2832	1.89	9.76	12.57	37
OH - NH2	315.0317	1.24	5.73	7.42	8

OH - NH	287.5930	1.20	6.06	6.78	6
OH - SH	38.9426	0.00	0.00	0.00	1
OH - COOH	147.0768	0.96	4.63	4.63	2
OH - EtherO	136.1517	1.79	8.58	10.29	52
OH - Epox	226.9612	0.00	0.00	0.00	1
OH - Ester	212.3077	2.49	11.46	13.18	18
OH - Ketone	46.7707	1.67	7.83	8.67	8
OH - Teth	-73.9896	0.38	1.99	2.26	4
OH - CN	306.7565	1.31	5.87	6.36	3
OH - AO	435.7842	0.00	0.00	0.00	1
OH - AN6	1333.3889	0.00	0.00	0.00	1
OH(a) - OH(a)	288.3543	1.05	5.57	6.09	4
OH(a) - NH2	795.5725	0.00	0.00	0.00	1
OH(a) - EtherO	130.0953	2.59	13.39	15.21	10
OH(a) - Ester	-1175.6769	0.00	0.00	0.00	1
OH(a) - Alde	36.1636	0.00	0.00	0.00	1
OH(a) - Nitro	-1049.0906	1.05	5.22	5.56	3
OH(a) - AN6	-617.7111	2.46	13.16	14.44	3
NH2 - NH2	174.2465	0.94	4.19	5.09	15
NH2 - NH	510.9788	0.98	5.05	5.46	4
NH2 - EtherO	124.8749	1.17	5.83	7.12	10
NH2 - Ester	188.4529	1.77	9.79	10.42	3
NH2 - Teth	-555.2077	0.78	4.00	4.00	2
NH2 - Nitro	666.8005	1.58	8.97	8.97	2
NH2 - AO	395.8822	0.00	0.00	0.00	1
NH2 - AN6	30.6550	1.74	8.51	8.51	2
NH - NH	240.0772	0.55	2.68	3.76	6
NH - EtherO	103.1723	0.56	2.36	2.65	6
NH - Ester	327.4470	0.00	0.00	0.00	1

NH - Ketone	-213.7974	0.00	0.00	0.00	1
NH - AN6	757.1224	0.00	0.00	0.00	1
SH - SH	218.1185	0.88	4.35	5.66	10
SH - Ester	502.0329	0.00	0.00	0.00	1
COOH - COOH	117.6044	1.46	8.38	10.28	4
COOH - EtherO	615.1998	1.56	8.74	12.66	6
COOH - Ester	-182.6418	2.17	12.14	12.15	2
COOH - Ketone	-55.6235	4.62	22.27	22.34	2
OCN - OCN	-362.3986	0.75	4.33	4.75	3
EtherO - EtherO	92.5659	1.70	7.72	10.43	185
EtherO - Epox	178.0583	1.93	8.99	9.78	3
EtherO - Ester	323.9389	1.77	8.56	11.81	25
EtherO - Ketone	16.3497	1.43	7.13	8.25	10
EtherO - Alde	17.5661	2.51	10.80	12.79	12
EtherO - Teth	393.8214	1.58	7.73	7.82	4
EtherO - Nitro	966.9443	0.71	3.89	3.89	2
EtherO - CN	293.4853	0.57	2.75	3.10	3
EtherO - AO	329.1098	0.34	1.35	1.35	2
Epox - Epox	1007.8569	0.00	0.00	0.00	1
Epox - Alde	164.1339	0.00	0.00	0.00	1
Ester - Ester	433.0853	2.01	10.10	12.62	69
Ester - Ketone	23.2513	2.42	11.50	13.86	25
Ester - Nitro	-205.2175	0.00	0.00	0.00	1
Ester - CN	517.3675	2.67	13.33	13.89	6
Ester - AO	708.4673	0.97	4.42	5.31	4
Ketone - Ketone	-303.6503	2.48	12.05	14.08	10
Ketone - Alde	-391.2766	0.00	0.00	0.00	1
Ketone - AtS	380.5159	0.45	2.29	2.45	3
Ketone - CN	-574.5412	0.00	0.00	0.00	1

Ketone - AO	179.4960	0.31	1.48	1.73	3
Ketone - AN6	123.6293	0.84	4.22	4.61	3
Alde - Alde	583.2711	0.00	0.00	0.00	1
Alde - Nitro	145.1830	0.06	0.33	0.33	2
Alde - AtS	396.0294	0.00	0.00	0.00	1
Alde - AO	674.7480	0.00	0.00	0.00	1
Teth - Teth	-11.7870	1.65	7.86	9.42	7
Nitro - Nitro	62.5023	0.91	5.42	5.96	3
AtS - CN	-102.6614	0.00	0.00	0.00	1
AtS - AN5	-350.3396	2.09	9.19	10.11	10
CN - AN6	-373.5012	0.00	0.00	0.00	1
AO - AN5	-890.1562	0.00	0.00	0.00	1
AN6 - AN6	-274.0201	1.82	7.42	8.17	3

Appendix E

Normal Boiling Point File Reconstruction

(All files are referenced to the CD on the back cover of this thesis. Reference to 'Tb-method cd – version.xls' file is assumed, unless otherwise stated.)

1) The ink files for the first-order structural groups ('Tb-method cd-version.ink') and corrections ('corrections-cd-version.ink') are prepared.
2) The respective automatic fragmentation procedures are generated for the ink files. The csv files ('Tb-method cd-version.csv'and 'corrections cd-version.csv') are then saved.
3) The startup file 'Tb-empty cd-version.xls' is opened. The 'global' module is then modified to match current data settings. The 'global' declaration declares all data settings global to all modules. Thus, the global routine is the control routine for all modules. The routine 'Start_from_scratch' from the 'start' module is then run. This routine creates the new Excel file ('Tb-method cd-version.xls') with a 'data' worksheet (Figure E1 and E2) and merges all structural group information from the 'csv' files.
4) All updated modules are then copied to the new file. The corrections are now merged, by first copying the worksheet from the csv file, and then by running the 'merge_corr' routine.
5) The 'worksheet_format' routine is then run. This routine performs most of the operations for the construction of the file. This includes (routines in parenthesis):

   5.1) Importing group information from the ink file ('loadgroups') and from data files ('retrieve_nbp' and 'retrieve_prop').

   5.2) Creation of auto-filters ('A_filter_main') and then removing all components with errors or no experimental boiling points ('A_filter_error_remov'). Deleting these components will speed up the performance, since only about 3000 components out of the 17000 components fragmented, are used.

5.3) The 'data' worksheet is then formatted ('My_Format'), for example, column width of structural groups is changed to a smaller size. This makes it easier to view.

5.4) The steric and isomer correction is then generated ('calc_icor_steric').

5.5) The sum of group contributions (Equation 4-1) and number of atoms for each component is then calculated ('group_calc'). This routine only calculates for groups where a frequency exists. This routine must be run before the group interaction metalanguage, since it is dependent on the number of atoms of the molecule.

5.6) The custom views are then created ('Groups_on' and 'Groups_off').

5.7) The normal boiling point estimation of the available group contribution methods for each component are then imported into the worksheet ('method_estimate'). This uses an OLE server (Rechenmodul.CalculationsMethods).

6) Currently, the ink file does not allow fragmentation for a short chain group for alcohols. This is done by a routine ('get_short_OH' in 'stuff' module).

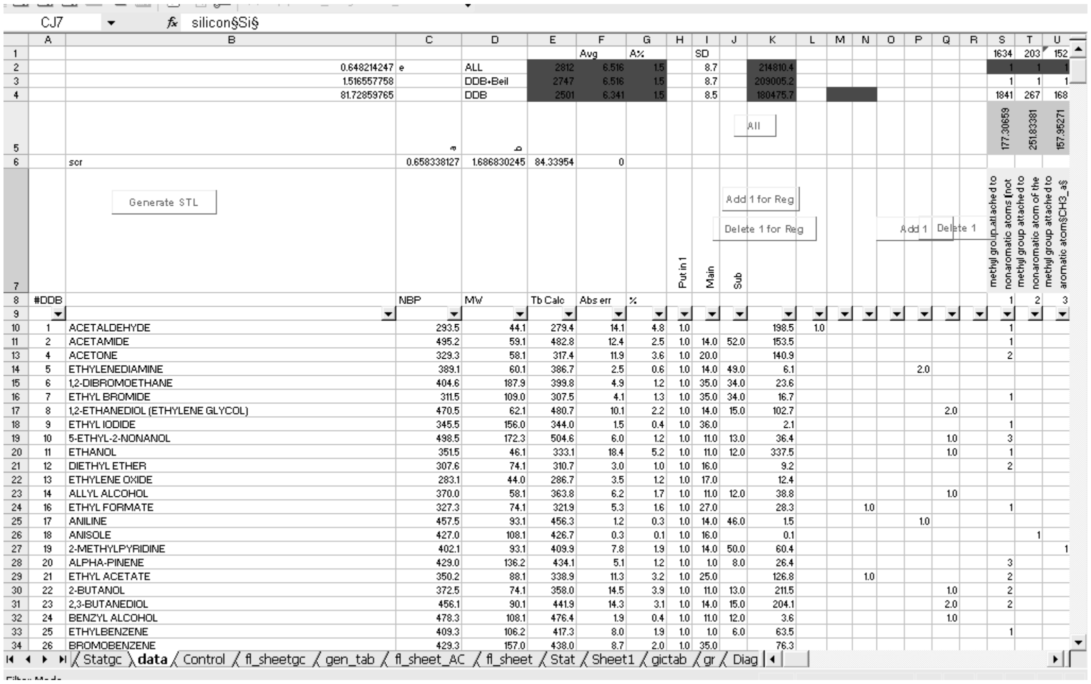
7) The worksheets 'Control', 'fl_sheet', 'fl_mf' and 'R_data' must be copied and updated. The 'Control' worksheet contains input for the regression. The 'fl_sheet' and 'fl_mf' worksheets are the filter metalanguage (filter language) and interaction language, respectively. Two worksheets, 'Stat' and 'mstat', must be created for output of the above two metalanguages. The 'R_data' worksheet contains components that must be removed from the worksheet. This is done by a routine ('remov_bad_data' in 'start' module).

8) The interaction language must then be run from the 'fl_mf' worksheet. This can be done by pressing the 'run' button on the worksheet. The reference module to this metalanguage is 'M_filter_language'. After the group interaction parameters are generated, the 'group_calc' routine must be run to generate the sum of group contributions. This can be easily done by the 'Re-generate sum in col II and IJ' button in the 'data' worksheet (Figure E2).

9) The starting values for the regression (for the non-linear parameters) must be set in 'row 6' (Figure E1). The components chosen for the regression must be checked by adding a value of '1' in 'column H' (Figure E1). If a new model equation is chosen, then the equation must be created in the 'AUX' function in the regression module. The regression can now be run by pressing the 'run' button in the 'Control' worksheet.

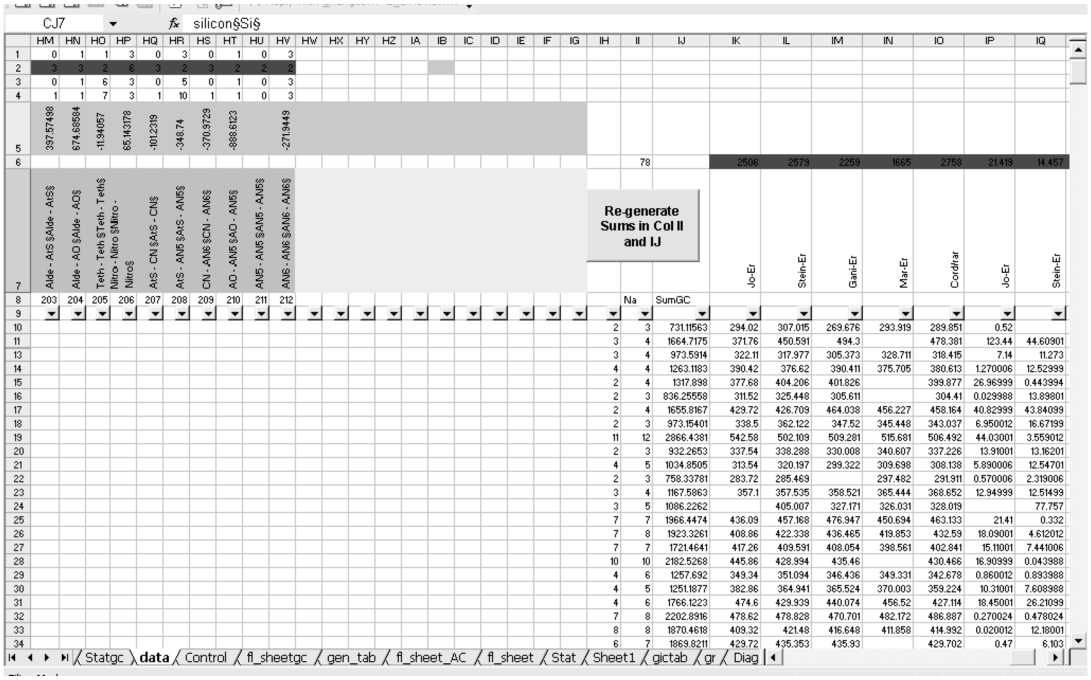
10) The file is now ready for analysis. The filter language can now be run by pressing the 'run' button in the 'fl_sheet' worksheet. This will generate a statistical analysis of

the functional groups in the 'Stat' worksheet.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1						Avg	A%		SD										1634	203	152
2		0.648214247	e	ALL	2812	6.516	1.5		8.7		214810.4								1	1	1
3		1.516557758		DDB+Beil	2747	6.516	1.5		8.7		209005.2								1	1	1
4		81.72859765		DDB	2501	6.341	1.5		8.5		180475.7								1841	267	168
5			a	b															177.30659	251.83381	157.95271
6		scr	0.658338127	1.686830245	84.33954	0															
7								Put in 1	Main	Sub									methyl group attached to non-aromatic atoms (not	methyl group attached to non-aromatic atom of the	methyl group attached to aromatic atom$CH3_a$
8	#DDB		NBP	MW	Tb Calc	Abs err	%												1	2	3
9																					
10	1	ACETALDEHYDE	293.5	44.1	279.4	14.1	4.8	1.0			198.5	1.0							1		
11	2	ACETAMIDE	495.2	59.1	482.8	12.4	2.5	1.0	14.0	52.0	153.5								1		
13	4	ACETONE	329.3	58.1	317.4	11.9	3.6	1.0	20.0		140.9								2		
14	5	ETHYLENEDIAMINE	389.1	60.1	386.7	2.5	0.6	1.0	14.0	49.0	6.1					2.0					
15	6	1,2-DIBROMOETHANE	404.6	187.9	399.8	4.9	1.2	1.0	35.0	34.0	23.6										
16	7	ETHYL BROMIDE	311.5	109.0	307.5	4.1	1.3	1.0	35.0	34.0	16.7								1		
17	8	1,2-ETHANEDIOL (ETHYLENE GLYCOL)	470.5	62.1	480.7	10.1	2.2	1.0	14.0	15.0	102.7						2.0				
18	9	ETHYL IODIDE	345.5	156.0	344.0	1.5	0.4	1.0	36.0		2.1								1		
19	10	5-ETHYL-2-NONANOL	498.5	172.3	504.6	6.0	1.2	1.0	11.0	13.0	36.4						1.0		3		
20	11	ETHANOL	351.5	46.1	333.1	18.4	5.2	1.0	11.0	12.0	337.5						1.0		1		
21	12	DIETHYL ETHER	307.6	74.1	310.7	3.0	1.0	1.0	16.0		9.2								2		
22	13	ETHYLENE OXIDE	283.1	44.0	286.7	3.5	1.2	1.0	17.0		12.4										
23	14	ALLYL ALCOHOL	370.0	58.1	363.8	6.2	1.7	1.0	11.0	12.0	38.8						1.0				
24	16	ETHYL FORMATE	327.3	74.1	321.9	5.3	1.6	1.0	27.0		28.3			1.0					1		
25	17	ANILINE	457.5	93.1	456.3	1.2	0.3	1.0	14.0	46.0	1.5					1.0					
26	18	ANISOLE	427.0	108.1	426.7	0.3	0.1	1.0	16.0		0.1									1	
27	19	2-METHYLPYRIDINE	402.1	93.1	409.9	7.8	1.9	1.0	14.0	50.0	60.4										1
28	20	ALPHA-PINENE	429.0	136.2	434.1	5.1	1.2	1.0	1.0	8.0	26.4								3		
29	21	ETHYL ACETATE	350.2	88.1	338.9	11.3	3.2	1.0	25.0		126.8			1.0					2		
30	22	2-BUTANOL	372.5	74.1	358.0	14.5	3.9	1.0	11.0	13.0	211.5						1.0		2		
31	23	2,3-BUTANEDIOL	456.1	90.1	441.9	14.3	3.1	1.0	14.0	15.0	204.1						2.0		2		
32	24	BENZYL ALCOHOL	478.3	108.1	476.4	1.9	0.4	1.0	11.0	12.0	3.6						1.0				
33	25	ETHYLBENZENE	409.3	106.2	417.3	8.0	1.9	1.0	1.0	6.0	63.5								1		
34	26	BROMOBENZENE	429.3	157.0	438.0	8.7	2.0	1.0	35.0		76.3										

Figure E-1: MS-Excel 'Data' worksheet showing columns A-U



	HM	HN	HO	HP	HQ	HR	HS	HT	HU	HV	HW	HX	HY	HZ	IA	IB	IC	ID	IE	IF	IG	IH	II	IJ	IK	IL	IM	IN	IO	IP	IQ
1	0	1	1	3	0	3	0	1	0	3																					
2	3	3	2	6	3	2	3	2	2	2																					
3	0	1	6	3	0	5	0	1	0	3																					
4	1	1	7	3	1	10	1	1	0	3																					
5	397.57498	674.68584	-11.94057	65.143178	-101.2319	-348.74	-370.9729	-888.6123		-271.9449																					
6																							78		2506	2579	2259	1665	2758	21.419	14.457
7	Alde - AtS $Alde - AtS$	Alde - AO $Alde - AO$	Teth - Teth $Teth - Teth$	Nitro - Nitro $Nitro - Nitro$	AtS - CN $AtS - CN$	AtS - AN5 $AtS - AN5$	CN - AN6 $CN - AN6$	AO - AN5 $AO - AN5$	AN5 - AN5 $AN5 - AN5$	AN6 - AN6 $AN6 - AN6$															Jo-Er	Stein-Er	Gani-Er	Mar-Er	Cordfrar	Jo-Er	Stein-Er
8	203	204	205	206	207	208	209	210	211	212													Na	SumGC							
9																															
10																						2	3	731.11563	294.02	307.015	269.676	293.919	289.851	0.52	
11																						3	4	1664.7175	371.76	450.591	494.3		478.381	123.44	44.60901
13																						3	4	973.5914	322.11	317.977	305.373	328.711	318.415	7.14	11.273
14																						4	4	1263.1183	390.42	376.62	390.411	375.705	380.613	1.270006	12.52999
15																						2	4	1317.898	377.68	404.206	401.826		399.877	26.96999	0.443994
16																						2	3	836.25558	311.52	325.448	305.611		304.41	0.029988	13.89801
17																						2	4	1655.8167	429.72	426.709	464.038	456.227	458.164	40.82999	43.84099
18																						2	3	973.15401	338.5	362.122	347.52	345.448	343.037	6.950012	16.67199
19																						11	12	2866.4381	542.58	502.109	509.281	515.681	506.492	44.03001	3.559012
20																						2	3	932.2653	337.54	338.288	330.008	340.607	337.226	13.91001	13.16201
21																						4	5	1034.8505	313.54	320.197	299.322	309.698	308.138	5.890006	12.54701
22																						2	3	758.33781	283.72	285.469		297.482	291.911	0.570006	2.319006
23																						3	4	1167.5863	357.1	357.535	358.521	365.444	368.652	12.94999	12.51499
24																						3	5	1086.2262		405.007	327.171	326.031	328.019		77.757
25																						7	7	1966.4474	436.09	457.168	476.947	450.694	463.133	21.41	0.332
26																						7	8	1923.3261	408.86	422.338	436.465	419.853	432.59	18.09001	4.612012
27																						7	7	1721.4641	417.26	409.591	408.054	398.561	402.841	15.11001	7.441006
28																						10	10	2182.5268	445.86	428.994	435.46		430.466	16.90999	0.043988
29																						4	6	1257.692	349.34	351.094	346.436	349.331	342.678	0.860012	0.893988
30																						4	5	1251.1877	382.86	364.941	365.524	370.003	359.224	10.31001	7.608988
31																						4	6	1766.1223	474.6	429.939	440.074	456.52	427.114	18.45001	26.21099
32																						7	8	2202.8916	478.62	478.828	470.701	482.172	486.887	0.270024	0.478024
33																						8	8	1870.4618	409.32	421.48	416.648	411.858	414.992	0.020012	12.18001
34																						6	7	1869.8211	429.72	435.353	435.93		429.702	0.47	6.103

Figure E-2: MS-Excel 'Data' worksheet showing columns HM-IQ

Appendix F

Filter Program Description

The development of the filter program involves the definition of commands and its properties. Figure F-1 presents an example of the filter program for primary amines, including the worksheets row number and column alphabet. The commands are defined as follows (column letters in parenthesis):

Command: goto (A)

Properties: Worksheets row number (B)

Description: The 'goto' command is located in the first line of the filter program. It is used for the case of when the program needs to only be run for a specific filter setting. For the case of 'Primary Amines', as in the example below, only the worksheet row number needs to be inserted in the properties cell. If the properties cell is empty, then the program starts from the beginning.

Command: filter (A)

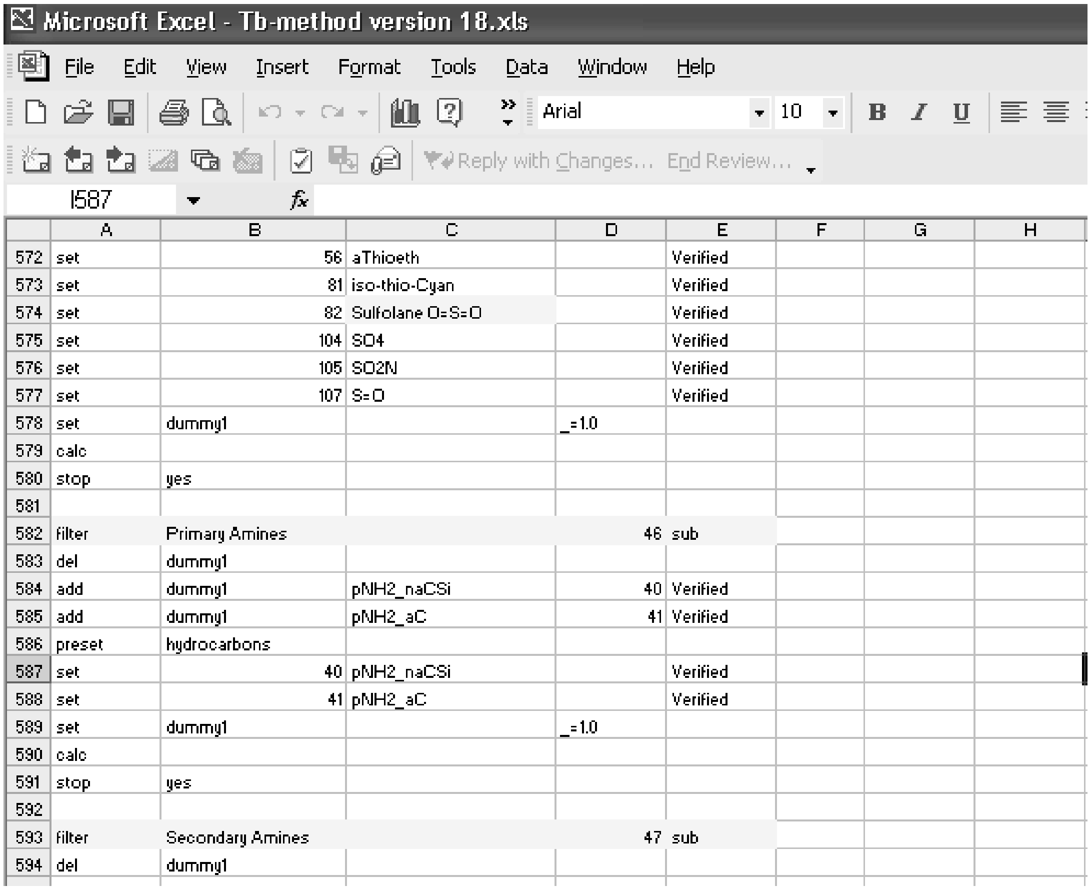
Properties: Filter setting name (B), Identification (ID) number (D), and main or sub-group.

Description: The 'filter' command indicates the start of a filter setting. It then stores the specific information about the filter setting, the name, ID and whether the group is part of a main group, sub-group or none. The latter distinctions are used to differentiate between similar functional groups. For example, a main group can be hydrocarbons, while a sub group can be n-alkanes. The main and sub-group also represent columns in the data worksheet. These columns store the group ID for each component. The storage of these ID numbers is a special case of object-oriented programming which prevents the ID numbers from being written into hidden cells. This makes it easier to call the setting, by just filtering the group ID in the worksheet.

Command: del (A)

Properties: dummy1 (B)

Description: Each auto-filter can filter a column by a custom setting (for example, cells with the value of 1), blanks (filters column cells which are empty) or non-blanks (filters column cells which are not empty). However, it is not possible to filter any of the above settings for more than one group. For example, alcohols are represented by 5 different structural groups. In order to generate alcohol components, the frequencies of each group needs to be added to an empty column, and this column now has to be filtered. Thus, a dummy column is used for this purpose. The 'del' command clears the contents of the 'dummy1' column.



	A	B	C	D	E	F	G	H
572	set	56	aThioeth		Verified			
573	set	81	iso-thio-Cyan		Verified			
574	set	82	Sulfolane O=S=O		Verified			
575	set	104	SO4		Verified			
576	set	105	SO2N		Verified			
577	set	107	S=O		Verified			
578	set	dummy1		_=1.0				
579	calc							
580	stop	yes						
581								
582	filter	Primary Amines		46	sub			
583	del	dummy1						
584	add	dummy1	pNH2_naCSi	40	Verified			
585	add	dummy1	pNH2_aC	41	Verified			
586	preset	hydrocarbons						
587	set	40	pNH2_naCSi		Verified			
588	set	41	pNH2_aC		Verified			
589	set	dummy1		_=1.0				
590	calc							
591	stop	yes						
592								
593	filter	Secondary Amines		47	sub			
594	del	dummy1						

Figure F-1: MS-Excel Worksheet showing Filter Program for Primary Amines

Command: add (A)

Properties: **d**ummy1 (B), group code (C), group number (D), verification (E)

Description: The 'add' command adds the frequencies of the respective groups for a filter setting to the 'dummy1' column. In the case of the example above, primary amines is represented by two groups, group number 40 (row 584) and 41 (row 585). The group number identifies the column in which the group is stored, in the data worksheet. The frequencies of both groups are then added to the 'dummy1' column. Thus, the 'dummy1' column now contains the frequency of all the primary amine components.

The group code is a shortened version of the group name within a section sign symbol (§). For example, pNH2_naCSi (row 584) is identified as 'primary amines attached to non-aromatic Carbon or Silicon'. This code is used for verification of the group. If the group is found in the worksheet, according to the group number in the filter setting, then the line is printed 'verified' (E). If the group is not found, then the program stops at that point.

Command: preset (A)

Properties: preset name (B)

Description: MS-Excel has a 'Custom View' (Section 4.3.1) setting which creates and stores different views of a worksheet. The custom view in only used when dealing with a filter setting common to most of the filter settings. For example, since mono-functional compounds uses hydrocarbons as the backbone, the hydrocarbon filter setting is common to all the functional groups. The hydrocarbon custom view turns 'off' the specified hydrocarbon groups, thus producing only hydrocarbon components. As nomenclature, the term 'on' refers to a filter with a custom, blank or non-blank setting and the term 'off' to a filter with no setting. As default, all the filters are set to a blank setting, or are turned 'on'. Thus, the advantage of the custom view is that its saves a large amount of time, since the filter program will have to generate the view, by turning 'on' or 'off' the column filters.

Command: set (A)

Properties: group number or dummy column (B), group code (C), column setting (D), verification (E)

Description: The description of the group number, group code and verification is the same as applied to the 'add' command. However, if the set command refers to a dummy column, then there is no group code and verification, since the 'dummy' column is already set in the 'add' command. The column setting sets the filter of that column to the specified setting. For the example above (Row 587), the cell is empty. This is a default setting to turn the filter 'off'.

Command: calc (A)

Description: The 'calc' command writes all the gathered and generated information into a statistic worksheet. The gathered information is settings collected in the filter setting, such as the filter name and ID number. The generated information is the information calculated in the worksheet, such as the average absolute deviation and number of components for the proposed and available methods. Other information can also be easily programmed. All the statistics are now written into the row corresponding to the ID number in a statistics worksheet.

Command: store preset (A)

Properties: preset name (B)

Description: The 'store preset' command creates a preset name, for example hydrocarbons, into custom views. When the name is created, the view would be exactly as the worksheet is presented.

Command: stop (A)

Properties: Boolean value (B)

Description: In the event of running just a single filter setting, the command 'stop' is used. The Boolean value 'yes' is used to stop the filter program at that particular point. Any other value will be a default to continue onto the next filter. For the example above, the 'goto' command must first be set to the row number of the filter (row 582) and the 'stop' command must be set to 'yes'.

Command: end (A)

Description: The 'end' command ends the filter program entirely. If there is 'end' command is omitted, then the default row number for the program to end is 32000.

Appendix G

Group Interaction Metalanguage Description

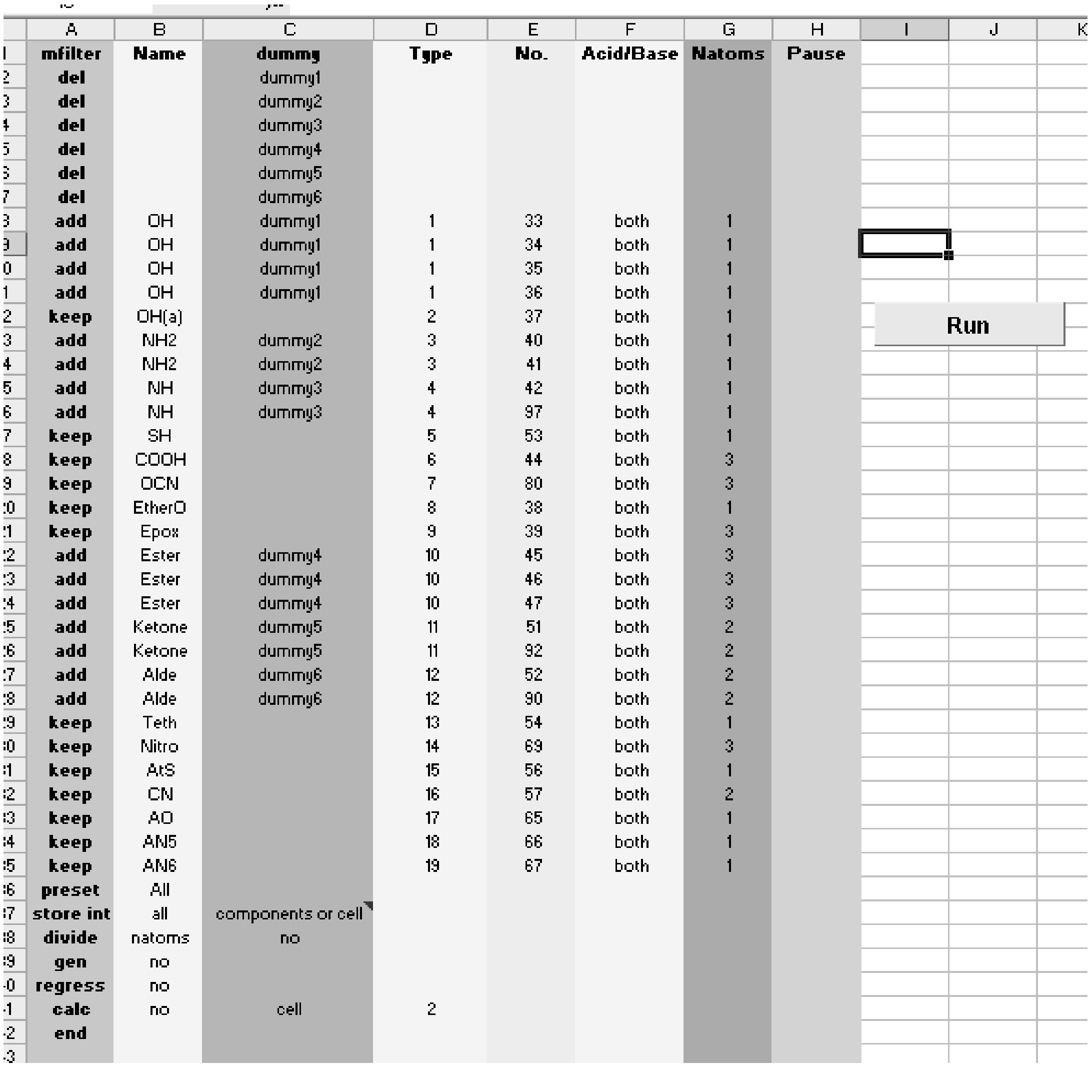
The development of the metalanguage involves the definition of commands and its properties. Figure G-1 presents the entire interface of the metalanguage, including the worksheets row number and column alphabet. The commands are defined as follows (Column letters in parenthesis):

Command: mfilter (A)
Description: The 'mfilter' command indicates the start of the metalanguage.

Command: del (A)
Properties: dummy column (C)
Description: The 'del' command is quite similar to the same command defined in the filter program. The only change is that different dummy columns are now used to store different combination of groups. In this case, only 6 columns are required to store these groups. Thus, the columns are labelled 'dummy1' … 'dummy6'.

Command: add (A)
Properties: name (B), dummy column (C), interaction ID (D), group ID (E), acid/base (F), number of atoms (G), Pause (H)
Description: The 'add' command is again similar to the same command defined in the filter program. The command is also used to store the description of the group. Firstly, the name of the group is stored, which is referenced to the name of the interaction parameter. For example, the group name 'OH' is part of the interaction group 'OH-NH', which is stored in the 'data' worksheet. The dummy column is as described previously. The interaction ID is an identification number for the program only. The frequency of the respective group is stored in matrix comprising of the component DDB number and interaction ID. This matrix is the input to the interaction frequency calculation program. The group ID, as before, is reference to the group

column in the 'data' worksheet. The acid/base property was used in the earlier development of the program, which designates whether the group is an acid or base. The concept of group interaction has been modified, and this property is no longer implemented, a default value of 'both' is used. The next property is the number of atoms of the group, excluding hydrogen. The 'pause' property is used to pause the program at that particular point, with a value of 'yes'. This is generally used for debugging.



A	B	C	D	E	F	G	H
mfilter	**Name**	**dummy**	**Type**	**No.**	**Acid/Base**	**Natoms**	**Pause**
del		dummy1					
del		dummy2					
del		dummy3					
del		dummy4					
del		dummy5					
del		dummy6					
add	OH	dummy1	1	33	both	1	
add	OH	dummy1	1	34	both	1	
add	OH	dummy1	1	35	both	1	
add	OH	dummy1	1	36	both	1	
keep	OH(a)		2	37	both	1	
add	NH2	dummy2	3	40	both	1	
add	NH2	dummy2	3	41	both	1	
add	NH	dummy3	4	42	both	1	
add	NH	dummy3	4	97	both	1	
keep	SH		5	53	both	1	
keep	COOH		6	44	both	3	
keep	OCN		7	80	both	3	
keep	EtherO		8	38	both	1	
keep	Epox		9	39	both	3	
add	Ester	dummy4	10	45	both	3	
add	Ester	dummy4	10	46	both	3	
add	Ester	dummy4	10	47	both	3	
add	Ketone	dummy5	11	51	both	2	
add	Ketone	dummy5	11	92	both	2	
add	Alde	dummy6	12	52	both	2	
add	Alde	dummy6	12	90	both	2	
keep	Teth		13	54	both	1	
keep	Nitro		14	69	both	3	
keep	AtS		15	56	both	1	
keep	CN		16	57	both	2	
keep	AO		17	65	both	1	
keep	AN5		18	66	both	1	
keep	AN6		19	67	both	1	
preset	All						
store int	all	components or cell					
divide	natoms	no					
gen	no						
regress	no						
calc	no	cell	2				
end							

Figure G-1: Screen shot of group interaction metalanguage interface.

Command: keep (A)

Properties: name (B), interaction ID (D), group ID (E), acid/base (F), number of atoms (G), Pause (H)

Description: The 'keep' command is similar to the 'add' command, involving only a single structural group. In this case, no dummy column is required. The properties are exactly the same as the 'add' command properties.

Command: preset (A)

Properties: preset name (B)

Description: The 'preset' command is exactly the same as the command defined in the filter program.

Command: store int (A)

Properties: components (B), cell number label (C), cell number (D)

Description: This is the most important command of the metalanguage. The 'store int' is where the group interaction frequency calculation is performed. The output is a three dimensional matrix comprising of the component DDB number and the two corners of an interaction parameter. For example, component 15, which has a frequency of 1 for interaction groups 1 and 4, is represented as (15, 1, 4) = 1. The command then prints the interaction parameters into the columns of the worksheet. It only prints parameters for which a frequency exists. The command also has a debugging tool to calculate a frequency of a specific component. In this case, the components property (B) cell is set to an empty cell, and the cell row number of the component is then inserted into the cell number property (D). This is used to verify a particular frequency of a component.

Command: divide (A)

Properties: Type of atoms (B), Boolean value (C)

Description: The 'divide' command divides the frequency of the parameter by the type of atoms. This requires the Boolean value (C) to be set to 'yes'. The type of atoms property involves two columns in the 'data' worksheet. The first is a fixed column containing the total number of atoms, excluding hydrogen. The second is a variable column comprising of different descriptions of the number of atoms, for example, number of carbon atoms. Thus, in this property cell, these columns are set.

Command: gen (A)

Properties: Boolean value (B)

Description: Since, group interaction parameter frequencies were added to the 'data' worksheet, the 'gen' command calculates the group contribution for each component in the worksheet.

Command: regress (A)

Properties: Boolean (B)

Description: The 'regress' command can perform the regression, including the new parameters.

Command: calc (A)

Properties: Boolean value (B), cell label (C), cell number (D)

Description: The 'calc' command is similar to the 'calc' command in the filter program. This command generates statistical results for each interaction parameter. The results are printed in 'mstat' worksheet starting from the cell number provided in the command property (D).

Command: end (A)

Description: The 'end' command ends the program.