University of Alberta

Prediction of the Molecular Structure of Ill-defined Hydrocarbons using Vibrational, ^{1}H and ^{13}C NMR Spectroscopy

by

Collins Obiosa-Maife

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

in

Chemical Engineering
Department of Chemical and Materials Engineering

Fall 2009
Edmonton, Alberta

Examining Committee Members

1. Dr. John M. Shaw (Supervisor)
2. Dr. William C. McCaffrey
3. Dr. Jeff Stryker

To those who persevere till their dreams see the light of day.
"*The desire accomplished is sweet to the soul*:" Proverbs 13:19

Abstract

This represents a proof-of-concept study of the appropriateness of vibrational and NMR spectroscopy for predicting the molecular structure of large molecules on the basis of a library of small molecules. Density Functional Theory (DFT) B3LYP/6-311G was used generate all spectra. 20 model compounds comprising two multiple-ringed polynuclear aromatic hydrocarbons (PAHs) connected by varying aliphatic chain-lengths were investigated. A least squares optimization algorithm was developed to determine the contribution of molecular subunits in the model compounds. ^{1}H and ^{13}C NMR spectroscopy failed to identify subunits unambiguously even with a constrained library. By contrast, IR and Raman results independently identified 40% and 65% respectively and jointly more than 80 % of the aromatic groups present; however, the aliphatic chain-length was poorly defined in general. IR and Raman spectroscopy are a suitable basis for spectral decomposition and should play a greater role in the identification of ringed subunits present in ill-defined hydrocarbons.

Acknowledgements

I arrived in Edmonton on a very cold January night; and I immediately realized that the elements would be a major challenge. I had to quickly adapt to the *unusually* short days and long wintry nights, the snow-covered roads, and the very hectic academic schedule. However, I found solace in a very supportive and generous supervisor, Dr. John M. Shaw. His many words of encouragement were sometimes therapeutic, and other times like an elixir; and the long hours of mentoring discussions provided the perfect *distraction* from the *hustle and bustle* of student life. Dr. Shaw's unique approach to conducting research, excellent guidance and patience provided the motivation for embarking on the onerous task of conducting the research documented in this thesis.

I am equally appreciative of the support and genial working atmosphere provided by past and present members of the *Petroleum Thermodynamics Research Group* including Bei, Vasek, Khan, Nima, Kasra, Keivan, Ardalan, Farshad, Jesus, Ala, Nafiseh, Merouane, Mildred and Pat.

I would also like to acknowledge the various levels of *technical* assistance I received from Jeff Sheremata, Imran, Hector, Kartik, Jaganath, Barath, Shankar, and Venkat; your ideas and suggestions are greatly appreciated. In addition, the support received from the administrative staff of the Department including Lily Laser and Marion Pritchard is also highly appreciated.

The sponsors of the NSERC Industrial Research Chair in Petroleum Thermodynamics are gratefully acknowledged: Alberta Energy Research Institute (AERI), ConocoPhillips Canada Inc., Imperial Oil Resources, Kellogg Brown & Root (KBR) Energy & Chemicals, NEXEN Inc., Shell Canada Ltd., Halliburton Energy Services, Total E & P Canada, and National Science and Engineering Research Council of Canada (NSERC).

Thanks to all my friends including the Akandes, Ogunsuyis, Oladirans, Ogundipes, Bukky Adegoroye, Ejike Ofuonye, Chikelu Agu, Kingsley Ezeagwula, Idowu Ohioze, Joshua Adetunji, Olumide Iwakun, Victor & Nike Ayo; your friendship is greatly appreciated. Special thanks to Saheed & Ronke Akonko for ensuring that I spent my first week in class rather than house-hunting. The encouragement and support received from the Nsiahs and Ogunwemimos is also acknowledged.

The encouragement and inspiration provided by my loving parents is graciously acknowledged; as well as the love and support of my Obiosa-Maife and Okoro families.

Finally, I would like to thank Amarachukwu, for her love, patience, companionship, encouragement and friendship while working on this degree. God bless you.

Table of Contents

List of Tables

List of Figures

Nomenclature

AA: Pure Aromatic ring or aliphatic chain compounds

a.u.: arbitrary units

API: American Petroleum Institute

CAA: Combination of Aromatic and Aliphatic compounds

CEOS: Cubic Equations of State

C, H, N, O, and S: Carbon, Hydrogen, Nitrogen, Oxygen, and Sulfur

DFT: Density Functional Theory

GC-EOS: Group Contribution Equations of State

IR: Infrared

Ni: Nickel

NMR: Nuclear Magnetic Resonance

PAH: Polynuclear Aromatic Hydrocarbon

PREOS: Peng-Robinson Equation of State

V: Vanadium

1 Introduction

1.1 Why study molecular structure?

From elementary physics and chemistry, we have come to appreciate the relationship between the structure of a system under investigation and its properties or functions. It doesn't matter how simple or complex the system is, most of its properties can be completely explained if there is a detailed description of its atomic, electronic, and molecular structure.

Medical science has been radically transformed with the discovery of the building blocks of life. The discovery of the molecular units of human life called DNA and RNA, and the subsequent mapping of the human genome has enabled scientists to better understand how some group of people develop immunity to certain diseases whereas others do not. In a recent article published in the *New England Journal of Medicine* [1], even though COPD (Chronic Obstructive Pulmonary Disease) has a prevalence of about 50% among heavy smokers, and is projected to be the third leading cause of death worldwide by 2020, some smokers still avoid COPD because innate and adaptive immunity is an integrated system where cells and molecules function in close connection.

The field of nanotechnology has benefited immensely from the manipulation of the molecular structure of materials to build functional systems. Nanotechnology enables the design of new materials at the atomic and molecular level. For instance, properly structured gold nanoparticles can be used to generate enough heat from light to treat cancer cells in humans, while carbon nanotubes which are 1/100,000 the diameter of human hair are so strong that they could be used to protect airplanes from lightning strikes and cool computer circuits.

From the foregoing, there is no doubt that understanding the molecular nature of matter is a key component to not only understanding their function and properties, but also improving their function and properties.

1.2 Crude Oil Molecular Structure

Crude oil or petroleum is a complex organic mixture of 10^5 to 10^6 different molecules [2]. It must frequently be processed to form products with hydrogen content different from that of the original feedstock; hence, the chemistry of the refining process focuses primarily on the production of better as well as marketable products [3]. Understanding the molecular composition of crude oil is therefore critical to selection of production and refining processes and refining conditions that translate into better economic value for the crude oil, and reduction in flow assurance problems resulting from solid deposits in pipelines, including wax, asphaltene, and hydrates during production [4].

Generally, crude oils are classified for economic value based on API gravity. Heavy oils have API gravities of 20 or less, extra or super-heavy oils 10 or less whereas a typical light crude oil has an API of about 36-38 [5]. Oil sands (or natural Bitumen) are sands saturated with heavy or extra-heavy oil and typically have API gravities between 6 and 12 [5]. The largest extra-heavy oil deposit is located in the Orinoco heavy oil belt, and contains about 90% of the world's reserves; while 81% of the world's bitumen reserves are located in Alberta, Canada, together, they contain 3,600 billion barrels of oil in place [6].

The depletion of conventional or light crude reserves precipitated by rising global demand for energy has created the need to refine and produce heavier crude oil sources such as bitumen which account for more than 50% of the world's reserve of fossil fuels [5, 7].

1.3 Thermodynamic Applications of Molecular Structure

The macroscopic properties of matter including heat capacity, viscosity, pressure, temperature, density, and the energy transformations that affect the changes in these properties constitutes branches of applied thermodynamics. The first and second laws of thermodynamics comprise the functional relationships that relate the energy transformations with these thermophysical properties.

With the increasing need to produce and refine heavier feedstocks, good predictions of thermophysical properties for reservoir simulation and surface facility design are very important in the petroleum industry. Key thermophysical properties include vapour pressure and phase behaviour more broadly, and heat capacity.

1.3.1 Group Contribution Equations of State (GC-EOS) Models

Efficient design of chemical processes including reservoir simulation requires a good understanding of phase behaviour and the ability to predict thermophysical properties. These properties can be obtained using an equation of state model; which are used to fit experimentally observed phase behaviour data using inputs based on measured pure component properties such as critical temperature and pressure. There are two major types of EOS models in use today namely the cubic equations of state (CEOS) and the group contribution equations of state (GC-EOS).

CEOS such as the Peng-Robinson (PREOS) are commonly used in reservoir simulation to predict phase behaviour of petroleum fluids. The vapour pressure

predictions must be accurate in order to provide a usable thermodynamic model. This is achieved by modifying the attractive parameter of the CEOS to include experimental or estimated vapour pressure data [8]. This approach works well for conventional oils due to the small amount of non-distillable fluid, typically about 5 wt% [8]. However, it cannot be applied to heavy oils and bitumen where roughly 60 wt% of the fluids are non-distillable [8].

GC-EOS models provide an alternative means of obtaining these critical properties for heavy oils and bitumen because they are based on the contribution of each molecular or structural groups present in the compound of interest. The original GC method of Coniglio *et al*. [9] developed for the Peng-Robinson CEOS was evaluated by Van Waeyenberghe [10]. For binary mixtures (hexane/hexadecane and benzene/ethylbenzene), the GC-EOS PR gave slightly better agreement with the experimental data than the critical-property-based PREOS. However, in a recent publication, Saber and Shaw [11] reported that the combination of the PREOS with the GC method developed by Marrero and Gani [12] gave better results than the PREOS. The authors based their results on model mixtures containing n-eicosane and n-decylbenzene. Computed density and bubble pressure values, along with other properties, are sensitive to the details of molecular structure, e.g.: anthracene vs. phenanthrene. Results obtained from this work are expected to facilitate better molecular definition, a key enabling knowledge base for phase modeling, and technology selection in the petroleum industry.

1.3.2 Heat Capacity Prediction

Heat capacity at either constant pressure or volume is the basis for energy models for simulators and is a tool for probing phase transitions of bitumen, residues, and asphaltenes because of the close relationship between molecular structure, vibration spectra, and heat capacity- particularly in the solid state [13]. Lastovka

et al. developed a generalized similarity concept for heat capacity [14, 15], and devised correlations based solely on the knowledge of the elemental composition for organic solids [16] and ideal gases [17]. The correlations apply to diverse pure organic compounds and complex mixtures containing C, H, N, O, and S, using universal coefficients. Quantum mechanical calculations played a key role in the development of the correlations but the details of molecular structure, beyond elemental composition have proven to be of secondary importance.

1.4 Heavy crude oil components

Organic material deposition, flocculation and plugging usually take place during heavy oil and bitumen production near the well bore and in the reservoir, during pipeline transportation, in the coke process during refining, during combustion, and in engine operations [18, 19]. The deposition is caused by the heavy components of the crude oil which self-associate, flocculate, precipitate or even polymerize at high temperatures and pressures [19]. Among the heavy petroleum components, asphaltenes seem to draw more attention than others, because of their intractable nature [20].

During petroleum refining, the asphaltene constituents are non-distillable and remain in the residua fuels as the distillable fractions are removed [18]. Asphaltenes agglomerate in crude oils and vacuum residues, and these aggregates are usually very poorly soluble in most organic solvents and petroleum liquids[19]. Since asphaltenes are structurally incompatible with many organic solvents, preventing or removing the deposits they form requires an understanding of their molecular arrangements [19].

The depletion of light crude oil reserves has created a need to refine heavier feedstocks including asphaltenes, heavy oils, vacuum residue, and bitumen. These ill-defined petroleum fractions, which typically possess aromatic carbon contents

exceeding 50% [13] cannot be studied by techniques such as gas chromatography, gas chromatography-mass spectrometry, or liquid chromatography, developed for hydrocarbons. Physical methods, such as infrared (IR), nuclear magnetic resonance (NMR), X-ray spectroscopy [21], and Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) [22] have proven more successful at unlocking their molecular structures. For example, using ultra-high FT-ICR MS, researchers at Florida State University have identified 20,000 distinct elemental compositions in petroleum crude oil [22].

1.5 Thesis Outline

The balance of this thesis is organized as follows. Chapter 2 contains a review of the current literature regarding molecular structure prediction methods and strategies. A brief discussion on the origins of petroleum is presented, as well as the chemical constituents of petroleum. The chapter closes with a brief discussion of the many-electron problem of quantum mechanics and how the Density Function Theory (DFT) provides a workable solution of this intractable problem.

In Chapter 3, the theoretical underpinnings of the methods employed to generate and analyze the results reported are presented. Thus a more detailed description of the DFT methods is discussed, and the novel spectral optimization algorithm is introduced and elaborated upon.

The results obtained from applying the methods in Chapter 3 are reported in Chapters 4 and 5. Chapter 4 begins with some more detailed information about the computational technique, and the rest of the chapter is a summary of the results from infrared and Raman analysis. Chapter 5 highlights the contrast between vibrational and Nuclear Magnetic Resonance (NMR) spectroscopy; the chapter discusses the results from analyzing NMR data.

In Chapter 6, the conclusions drawn from the results and discussions in the thesis are presented; and the future direction for the research is recommended

2 Literature Review

2.1 Introduction

Heavy oil and bitumen represent the most chemically complex variety of petroleum; and the elucidation of the molecular structure of their ill-defined, high molecular weight constituents poses a huge challenge to the petroleum chemist [23]. The focus of this work is linked to the development of a novel approach to molecular structure prediction for the ill-defined fraction of petroleum including asphaltenes, heavy oil, bitumen, and vacuum residua.

The first part of this review focuses on the various measurement techniques used to characterize heavy oil and bitumen fractions. The challenges encountered in processing heavier feedstocks can be attributed to the chemical character and amount of high-boiling constituents in the feedstock. Individual molecules in ill-defined hydrocarbons cannot be measured; hence a combination of techniques is used and molecular structures are proposed that are consistent with experimental data [24]. To obtain information about bulk properties, elemental analysis, average molecular weight, nuclear magnetic resonance (NMR), infrared (IR), and ultraviolet-visible (UV-Vis) measurements are required. Separation techniques, including short path distillation, high performance liquid chromatography (HPLC), gas chromatography (GC), and solvent solubility, SARA (Saturates, Aromatics, Resins, Asphaltenes) analysis, yield the amounts of various fractions [24].

This second part of this review focuses on the various molecular structure representations used to model and describe the thermophysical properties of the heavy oil and bitumen fractions of petroleum. The review focuses on the molecular representation of asphaltenes, which represent the heaviest molecular

weight fractions of heavy oil and bitumen, and the fraction posing the greatest industrial challenges throughout the production and refining process chain. There are currently two widely-held views on the molecular structures present in this fraction, namely, the archipelago and the pericondensed structures. Literature supporting these two perspectives is reviewed in detail.

2.2 Petroleum Origins and Chemistry

The first scientific theory on the origin of petroleum is attributed to 19th century scholars: Berthelot and Mendeleev [25-27]. Berthelot (1876) and Mendeleev (1878) both postulated that petroleum was formed deep within the earth's crust by inorganic material. This inorganic theory of the origins of petroleum dominated for a long time even though in 1863, the Canadian petroleum geologist T.S. Hunt had suggested the organic origin of petroleum [25, 27]. However, the first truly comprehensive theory of the organic origins of petroleum was put forward in the 1930's by German Chemist Alfred Treibs, who reported the similarity between the chemical structure of porphyrins (organic molecules found in petroleum) and chlorophyll pigments found in living organisms [25, 27]. In addition, the observation of optical activity in petroleum fractions, which the abiogenic hypothesis could not account for [28], provided conclusive evidence of the biogenic origins of petroleum.

Petroleum is therefore formed from the accumulation of hydrocarbons from living organisms; as well as the formation of hydrocarbons by the action of heat on biologically formed organic material [29]. The hydrocarbons synthesized directly from living organisms or from their molecules account for 10-20% of petroleum; while the remaining 80-90% involves the conversion of lipids, proteins, and carbohydrates of living material into the organic matter of sedimentary rocks [29]. Some of these organic compounds have carbon or skeletal structures that can be traced back to a living organism or source material, and are called biomarkers.

Biomarkers provide essential information about the geological history of oils and source rocks [30].

The different hydrocarbon groups found in petroleum including heavy oils and bitumen are the paraffins, naphthenes or cylco-alkanes, and aromatics. The paraffins or alkanes (C_nH_{2n+2}) are the second most common constituents of crude oil. They form both straight and branched-chains. The straight chain paraffins are called normal paraffins or *n*-paraffins. The terms saturated and aliphatic hydrocarbons are also used for this group. The naphthenes or cyclo-alkanes (C_nH_{2n}) are the most common molecular structures in petroleum, and are formed by joining carbon atoms in a ring. The average crude oil contains about 50% naphthenes, with the quantities increasing in the heavier fractions and decreasing in the lighter fractions [29]. The naphthenes and paraffins are both referred to as saturated hydrocarbons because all the available carbon bonds are saturated with hydrogen. The aromatic hydrocarbons contain at least one benzene ring. They rarely account for 15% of the entire crude oil, and tend to be concentrated in the heaviest fractions including heavy oil, bitumen, and vacuum residua; in which they make-up about 50% [29]. In addition to hydrocarbons, petroleum also contains heteroatoms such as sulfur (S), nitrogen (N), and oxygen (O).

2.3 Heavy oil and Bitumen Definitions

Natural bitumen (often called tar sands or oil sands) and heavy crude oil differ from light crudes by their high viscosity (resistance to flow) at reservoir conditions, high density (low API gravity), and significant contents of heteroatoms including nitrogen, oxygen, and sulfur compounds, and heavy-metal compounds [6].

They are the products of previously light crude oil that have lost their light-molecular-weight components through bacterial degradation, water-washing, and

evaporation [5, 6]. The loss of light fractions renders the oil heavy with a high proportion of asphaltic molecules, which largely determine the increase or decrease in the density and viscosity of the oil [31]. They resemble the residuum from the refining of light oil, and are more costly to produce and transport; requiring additional upgrading to reduce their viscosity and carbon content before they can be used as feedstock in a conventional refinery [6].

Heavy oils, bitumen and vacuum residua are the heaviest fractions of petroleum. They contain larger quantities of non-distillable hydrocarbons, which are characterized by their higher molecular weight, lower H/C ratio, higher fraction of aromatic carbon, higher S and N heteroatom content, and trace levels of Ni and V [32].

2.4 Heavy Oil and Bitumen Composition Analysis

Heavy oils and bitumen like other petroleum fractions consist of hydrocarbons, resins and asphaltenes [33]. However, compared to conventional or light crude oils, heavy oils and bitumen have higher molecular weight fractions including asphaltenes and resins [33]. This increase in molecular weight distributions also correlates with increasing boiling point thereby rendering distillation methods used to analyze light crudes ineffective in heavy oil and bitumen analysis [34]. The higher the boiling point of a fraction, the more difficult it is to analyze [34].

For instance, according to Altgelt and Boduszynski [34], the composition of naphthas, with a normal boiling range from the start of the distillation process to 130 °C, can be easily determined by gas chromatography (GC) alone; whereas the composition of heavy naphthas (boiling range 130 °C-220 °C) is more difficult, and requires the use of gas chromatography/mass spectrometry techniques

(GC/MS) to analyze. However, the non-distillable fraction or vacuum residue, boiling above 700 °C are by far the most difficult to analyze due to their limited volatility and solubility which greatly limit the use of conventional analytical techniques for their analysis.

Due to this complexity, it is often impossible, and usually unnecessary to determine the individual molecular constituents of the oil [34]; compositional analysis are usually conducted by fractionation into predefined chemical groups [35] .The isolation of the different fractions is facilitated by the difference in solubility of petroleum in paraffin solvents; the soluble fraction is called maltenes, while the insoluble fraction is called asphaltenes. As mentioned in Chapter 1, asphaltenes are an important component in understanding the structure of heavy oils and bitumen; and they will be the subject of the next section.

2.5 Asphaltene Chemistry and Molecular Structure

Asphaltenes are the most aromatic, highest molecular weight, and the most polar constituents of crude oil; and were the subject of a very early investigation by Boussingault in 1837 [36-38], who described them as the distillation residue of bitumen.

They precipitate from crude oil upon the addition of nonpolar, low-boiling hydrocarbon solvents such as *n*-pentane, and *n*-heptane [18, 36, 39]. The yield and quality of asphaltenes depends on how it is separated, the time before filtration, the pore size and material of the filter [2].

Table 2-1: Elemental Composition of Athabasca [40], Duri [39]and Boscan [39] Asphaltenes

Wt.%	C	H	N	O	S
Athabasca	79.9	8.3	1.2	2.8	7.8
Duri	87.35	8.22	1.47	2.07	0.39
Boscan	81.10	7.79	1.11	1.43	6.10

The molecular structure of petroleum asphaltenes is fundamental to understanding its colloidal, solubility, thermal and chemical properties[39] . The attempts to elucidate the molecular structure of asphaltenes have been hampered by the high values and broad distributions of molecular weights and the extreme diversity of structural elements within the constituent molecules[40]. The common features of the general structures proposed involve variable sizes of polycondensed aromatic nuclei, naphthenic rings, sulfide linkages, alkyl side chains, and bridges between rings [40].

Although the molecular building blocks of asphaltenes are known [37], the molecular structure representation of these fragments has been contentious and very ambiguous. T. F. Yen [41] postulated that the asphaltenes were pericondensed in nature based on NMR and X-ray results, whereas Strausz *et al.* contend that pericondensed aromatic units play a minor role in petroleum asphaltenes [37] based on the ruthenium-ions-catalyzed oxidation (RICO) of Athabasca bitumen asphaltenes. Recently, Zhao *et al.* [42] and Sheremata *et al.* [43] using a combination of NMR, elemental and molecular weight analysis, reported different molecular structural representations for the asphaltenes obtained from Athabasca bitumen. While Zhao and co-workers reported a pericondensed structure, Sheremata *et al.* report archipelago representations. In a paper published in 2009, Boek and co-workers [44] processing Sheremata *et al*.'s data were also able to report pericondensed molecular representations by making

changes to the molecular assembly mechanism. Differences were primarily linked to the nature of the substructures presumed to be present.

Hence, the ambiguity is not only restricted to the analytical methods employed, but also extends to the analysis of asphaltenes obtained from the same source. Furthermore, the average molecular structure proposed appears to indicate the relative importance researchers place in the type of analytical techniques used namely proton versus Carbon-13 NMR, or ion cyclotron spectroscopy and fluorescence [20].

However, a central unresolved question is which of these two structural models more closely approaches the average structure for these ill-defined petroleum fractions. Further, the phase behaviour of these fractions is a key element for the development of both refining and production processes [13]. Current models only capture a subset of the relevant physics and chemistry of the range of length scales asphaltenes intra-act and interact; hence, they are only correlative in nature [20].

2.5.1 The Pericondensed Model

The pericondensed model (Fig. 2-1) of asphaltenes is characterized by a large single aromatic core of carbon atoms from which aliphatic chains extend. This model is supported largely by NMR spectroscopy, X-ray diffraction, and fluorescence depolarization [43].

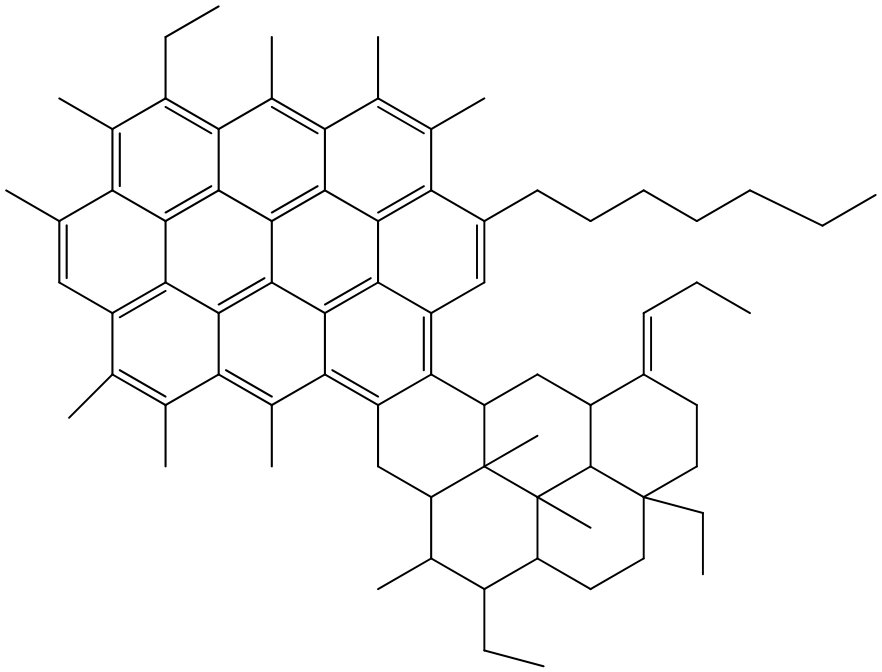


Figure 2-1: Pericondensed molecular structure representation (from Reference 46)

2.5.1.1 NMR Spectroscopy

Nuclear Magnetic Resonance (NMR) spectroscopy is a very useful technique for identifying and analyzing organic compounds, and it is based on the magnetic properties of atomic nuclear. The most common forms of NMR spectroscopy are those based on the nuclei of ^{1}H and ^{13}C. The ^{1}H gives information about the number of protons in an organic sample, while ^{13}C gives information about the carbon atoms present [45]. NMR measures aromatic and aliphatic carbon as well as hydrogen substitutions [34]. This technique can also detect the presence of various structural groups and provide information about their concentration [46].

Yen *et al.* [47] used proton-NMR to determine the fraction of saturated hydrocarbons in petroleum asphaltenes. The study also showed the presence of aromatic and naphthenic structures, and concluded that approximately 29 to 66 percent of the terminal carbons were in methyl or ethyl groups attached to aromatic centers. Sheremata et al. [43] used a combination of ^{1}H and ^{13}C, elemental analysis, and vapour pressure osmometry (VPO) in combination with

Monte Carlo simulations to elucidate the structure of Athabasca asphaltenes. They generated representative asphaltene molecules that were consistent with experimental data, and presented the first quantitative molecular representations of asphaltenes as archipelago structures.

However, both agree that NMR analysis alone cannot be used to predict the structure of complex molecular species such as asphaltenes because the aromatic structures present in asphaltenes exhibit very similar NMR signatures. Further, NMR results show the concentrations of various chemical groups including quaternary and protonated carbon structures, but not necessarily their spatial configurations. Thus it is possible to obtain multiple molecular representations from the same data set; hence the ambiguity of molecular representations of asphaltenes based on NMR data alone.

2.5.1.2 X-ray Diffraction

X-ray diffraction studies have been used to analyze the petroleum asphaltenes since the 1960s. Yen *et al*. [48, 49] used this technique to determine the aromaticity and crystallite parameters of eight petroleum asphaltenes, a petroleum resin, a gilsonite asphaltene, and an asphaltene obtained from a visbreaker tar. They concluded that asphaltenes were a condensed aromatic sheet, with naphthenic and aromatic branches. Shirokoff *et al*. [50] compared the crystallite and aromaticity parameters of asphaltenes from four Saudi Arabian crudes to the average structural parameters obtained from NMR; and arrived at the same conclusion.

However, Andersen *et al.* [51] suggest that the aromaticity values used to reach the conclusion by the previous authors were in disagreement with those expected from NMR and elemental analysis. While Altgelt and Boduszynski [34] contend that x-ray data can be misleading when used to translate geometric data into

structural information. In addition, they seem to suggest that the determination of aromaticity (the fraction of aromatic carbon over total carbon in a molecule) using x-ray data is inappropriate and arbitrary.

Hence, for the complete analysis of molecular structures, it is necessary to apply all available spectroscopic methods .[45]

2.5.1.3 Fluorescence Spectroscopy

Fluorescence spectroscopy methods use the optical characteristics of asphaltenes to predict their molecular weight and structures. Groenzin *et al.* [52] used fluorescence depolarization techniques to measure the rotational correlation times of individual asphaltene molecules. Their findings suggest that the controversy surrounding the molecular weight and structure of asphaltenes is due to their tendency to aggregate at very low concentrations, 0.06g/L. They propose that analytical techniques such as vapour pressure osmometry (VPO) which requires large concentrations of asphaltenes generally measure large molecular weights due to asphaltene aggregation. Hence, it is their opinion that asphaltenes are monomeric, consisting of a single large aromatic core surrounded by aliphatic chains corresponding to the low molecular weights (~ 750 g/mol) observed in their studies. In a related publication, Andrews et al. [53] measured the translational diffusion coefficients of asphaltene molecules at very low concentrations, before the onset of aggregation, and also concluded that asphaltene molecules are monomeric and not polymeric. They also conclude that self-association of asphaltenes are responsible for the diversity in molecular structures proposed for asphaltenes. Similar results and conclusions were drawn from using time-resolved fluorescence depolarization (TRFD) techniques on dilute asphaltene solutions [54].

However, Strausz *et al.* [55] in their critique of the use of fluorescence techniques on dilute asphaltene solutions posit that fluorescence decay time measurements on highly complex mixtures like asphaltenes can yield numerical information with no physical meaning, the measurements were made using inappropriate equipment and methods, and the yield of fluorescence per unit weight strongly depends on VPO molecular weight of the asphaltene; which monotonically falls off with increasing molecular weight of the separated asphaltene fractions, and the highest molecular weight fractions (>17 000 g/mol) barely emit.

In a rebuttal to the above observations, Mullins [56] argues that recent results using various analytical techniques support the asphaltene molecular weight and structure predicted using the TRFD technique, and not the results of Strausz *et al.* [55]. He contends that because the TRFD is a non-destructive analytical method, it is superior to the destructive methods employed by Strausz et al. However, he does admit that it is possible to have an asphaltene molecular fraction with more than one polynuclear aromatic hydrocarbon (PAH); even though for over a decade the TRFD results have suggested that there is a single PAH per asphaltene molecule as the dominant structure.

2.5.2 The Archipelago Model

The archipelago model (Fig. 2-2) as the name suggests represents asphaltenes as a collection of aromatic “islands” connected by aliphatic “bridges”. This model is supported by data from pyrolysis, oxidation, and thermal degradation [43].

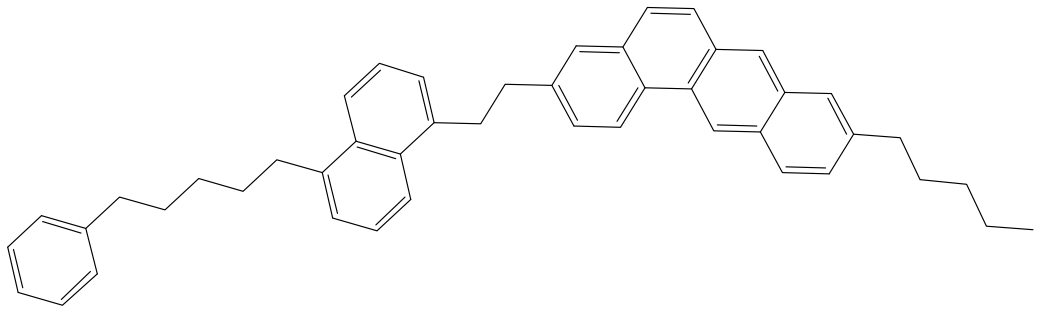


Figure 2-2: General archipelago structural representation (from Reference 46)

2.5.2.1 Pyrolysis, Chemical and thermal degradation

Strausz *et al*. [57] posit that the extent of aromatic condensation in asphaltenes is very low, hence the pericondensed model does not adequately describe the molecular structure of asphaltenes. In their paper, they point out that the infrared, NMR, and X-ray diffraction studies on asphaltenes conducted in the 1950's could not account for the presence of naphthenic structures in asphaltenes, which resulted in a predominance of aromatic and aliphatic structural components. Further, they argue that X-ray diffraction studies of Dickie and Yen [58] which suggested that asphaltenes were pericondensed, suffered from an "overestimation of the X-ray diffraction signal attributed to large aromatic disks"; thereby rendering the method and the predicted model unsuitable to asphaltenes.

They further argue that the products of mild thermolysis of Athabasca Bitumen asphaltene could not have been created by the break-up of large condensed aromatic clusters. Furthermore, they point out that the presence of copious amounts of biomarkers in asphaltenes indicated the presence of sulphides, sulphoxides and carboxylic acids which must be accounted for in any model of asphaltene molecular structure.

According to Gray [59] asphaltenes must have a variety of aromatic groups connected by bridges and substituted aliphatic groups, large molecular weight,

and have a large concentration of high boiling fractions (typically ~ 650C) to be consistent with experimental observations from cracking and pyrolysis. He suggests that an asphaltene molecule with a condensed aromatic core cannot give significant yields of volatile products in pyrolysis experiments, whereas an asphaltene molecule with a distribution of aromatic centers linked by aliphatic chains would produce a very wide range of products from methane to toluene-insoluble carbon residue.

Peng *et al*. [60] used a combination of nickel-bromide reduction of the carbon-sulfide bonds, basic hydrolysis of the ester bonds, and the boron tribomide cleavage of the carbon-oxygen bonds to study the chemolysis products of Athabasca asphaltenes, and concluded that the molecular structure of asphaltenes was characterized by core segments bound by acyclic sulfur linkages.

Finally, Jaffe *et al*. [61] also confirm that the archipelago model more closely describes the behaviour of asphaltenes under selective oxidation or pyrolysis. Hence, the archipelago molecular structure model is more consistent with the empirical chemolysis and thermolysis products of asphaltenes.

2.6 Molecular Representations

The chemical complexity and diversity of heavy petroleum fractions including asphaltenes make the analysis and modeling of their molecular structures very daunting. It is not possible to analyze and quantity the individual components or compounds or isomer lumps, because the number and complexity of the molecular structures grows rapidly with increasing molecular weight and boiling point [61, 62].

Average molecular structures have been used to represent vacuum residua (including asphaltenes). This approach uses correlations that rely on ^{1}H NMR data

to calculate average properties including aromaticity and degree of substitution [43, 44]. Although these structures are based on characterization information, they are inadequate for modeling purposes, because real conversion processes produce a wide variety of single-core molecules, which is inconsistent with a single starting molecule [61] . In addition, they suffer from limitations due to the complexity of the mixtures of heavy petroleum fractions.

More detailed molecular representations have been developed using stochastic and quantitative techniques. These techniques allow for a statistical representation of molecular structure of heavy petroleum feedstocks using analytical data as input.

2.6.1 "Average Molecule" Representation

According to Boduszynski [62], the complexity of petroleum mixtures increases rapidly with increasing boiling point due to the increasing number of atoms in each molecule and the vast number of potential structural arrangements or isomers. Hence, the characterization of high-boiling petroleum fractions such as vacuum residue in terms of individual components is practically impossible. Even a combination of the most powerful analytical tools including gas chromatography and mass spectrometry (GC/MS) is limited to relatively low-boiling petroleum fractions because of the immense number of structural isomers present in the high-boiling fractions rather than sample volatility.

Consequently, the molecular characterization of asphaltenes, heavy oils and vacuum residue has involved the "grouping" of compounds, and the concept of "average molecular structure"[62]. The concept of average molecular structure determination usually relies on the results of elemental analysis, ^{1}H and ^{13}C spectroscopy, and average molecular weight determination to derive the structural parameters of the average molecule [63].

2.6.1.1 Limitations of the Average Structure Concept

Boduszynski [63] contends that although the concept of the average molecule has been widely used to characterize heavy end petroleum fractions and coal-derived liquids from the late 1950's, it does not provide the required structural information for these complex mixtures.

He posits that heavy ends and vacuum residue are represented by a wide molecular weight distribution which also extends to relatively small molecules rather than consisting wholly of high molecular weight components. This wide distribution in molecular weights resulting from significant differences in volatility between the different chemical structures that make up these complex mixtures is the major limitation of the average structure model. The high concentration of polar, hetero-atom containing, polycyclic aromatic compound classes in the high boiling fractions and vacuum residua explains the wide molecular weight distributions of heavy ends.

He also points out that the use of vapour pressure osmometry (VPO) in average molecular weight determination of asphaltenes (with high concentrations of polar components) suffers from strong intermolecular associations. These intermolecular associations results in the over-estimation of the molecular weights of asphaltenes using VPO.

According to Petrakis and Allen, the analytical data on which the average structure is based are averaged over all the components in the mixture; and as a result the average structure is unlikely to be the dominant one, and may even be absent [64].

As a result of these limitations, and the number of molecules required to represent petroleum mixtures, more detailed schemes are required to better describe the molecular structure of heavy petroleum fractions

2.6.2 Quantitative Molecular Representation

The need to accurately represent the process chemistry of petroleum feedstocks as a means to obtaining good predictions on product properties has led to the development of more sophisticated modeling techniques. For complex hydrocarbon mixtures, it has been operationally expedient to use lumped kinetic schemes, where molecular components are generally grouped by boiling point and solubility classes ; these lumps are then represented in reaction models as if each were a single pseudo-component [65]. However, these lumps contain several molecules that are chemically dissimilar other than the definition of the lump to which they belong [65, 66]. The limitations of the lumped models motivated the development of a molecular-based approach with the goal of assembling a set of representative molecules with properties characteristic of a given feed [65].

Quantitative representations use a combination of advanced analytical characterization techniques including ^{1}H, ^{13}C NMR and mass spectrometry, and statistical modeling approach to transform structural attributes for heavy petroleum fractions into detailed molecular structure representations [65, 67].

2.6.2.1 Structure-Oriented Lumping

Quann and Jaffe [66] developed the structure-oriented lumping (SOL) for describing the composition, reactions and properties of complex hydrocarbons mixtures in 1992; and in 2005, the method was extended to incorporate vacuum residua by Jaffe et al [61] Structure-oriented lumping represents individual hydrocarbon molecules as vectors of 22 structural increments. The SOL increments (Fig. 2-3) consist of three aromatic increments, six naphthenic increments, a side-chain increment, and two isomeric descriptors: branches on side chains and methyl groups [61]. In addition, a hydrogen increment is used to represent unsaturation, two sulfur increments, three nitrogen and three oxygen

increments are used to describe heteroatomic hydrocarbons; while a biphenyl bridging increment is also specified.

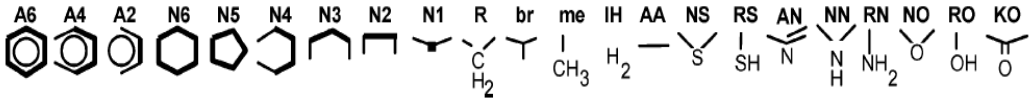
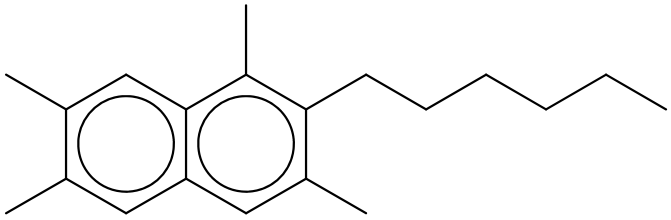


Figure 2-3: Depiction of the structure-oriented lumping (SOL) structural increments [61]

The SOL approach represented gas oil and lighter petroleum fractions in terms of homologous series of single-core molecules such as benzene and alkylated benzenes, and heavy fractions such as vacuum residua as multicore molecules comprising linked assemblies of single-core species [61]. The molecular description is achieved by representing each complex mixture as a table of vectors; where each row represents a molecule with each having a concentration in weight percent [61]. Fig. 2-4 shows a hydrocarbon structure in the SOL representation.



A8	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	1	0	0	0	0	0	0	0	10	0	4	0	0	0	0	0	0	0	0	0	0

Figure 2-4: An example of the SOL representation of a hydrocarbon structure (from reference 66)

Different molecules with the same set of structural groups (isomers) are lumped and represented by the same vector; however, this lumping is achieved at the molecular structure level [66]. The structure vector framework allows for the construction of rule-based reaction networks involving thousands of components

and many thousands of reactions. In addition, the multicore representation of vacuum residua allows for the direct calculation or modeling of their physical and chemical properties including elemental composition, molecular weight, boiling point and density [61].

2.6.2.2 Monte Carlo Methods

Neurock *et al*. [67] developed a Monte Carlo scheme that allowed for the stochastic construction of the molecules of heavy petroleum fractions such as vacuum residue and asphaltenes through random sampling of probability density functions (PDF's). The PDF provided the quantitative probability of finding the value of a given structural attribute [65]. The attributes, which include number of aromatic rings, number of naphthenic rings, and number and length of aliphatic sides chains, represent the key structural elements that provide the basis for the molecular description of the heavy petroleum fractions [67]; and their probabilities of occurrence in a feedstock were determined by using a combination of distillation and solubility measurements with a set of analytical measurements such as ^{1}H NMR, vapour pressure osmometry (VPO), and elemental analysis.

The four basic steps involved in the scheme were (1) the deduction of a chemical logic diagram, (2) the compilation of structural cumulative probability distributions, (3) stochastic sampling of each distribution, and (4) the construction of molecular species.

This four-step approach was applied to an offshore California asphaltene sample. The asphaltene molecule was described in terms of unit sheets containing aromatic and naphthenic rings, the aliphatic chains attached to each unit sheet, and respective lengths of each chain. PDF's for each of these attributes were determined by assigning either a Gaussian or gamma function. The complete

molecular representation of the asphaltene feedstock was determined by stochastically sampling 10,000 different molecules.

This method was extended by Sheremata *et al*. to include structural data from quantitative ^{13}C NMR [43]. Sheremata's work was the first quantitative molecular representation of asphaltenes as archipelago structures. An asphaltene sample obtained from Athabasca vacuum residue was used in this study. The gamma distribution was chosen for the PDF's. The aromatic ring groups used in the study are presented in Figure 2-5.

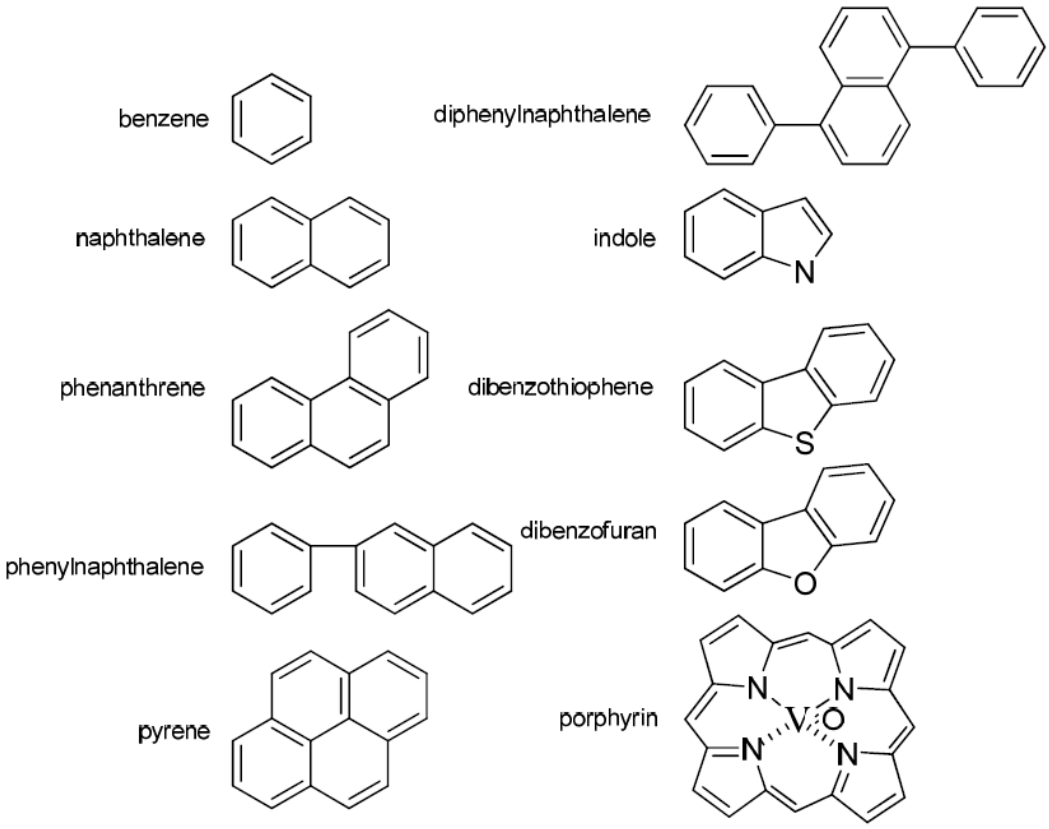


Figure 2-5: Aromatic ring groups used in Sheremata et al.'s [46] asphaltene representation

Half of the naphthenic carbons were assumed to be associated with aromatic rings, and the other half with aliphatic carbons [43]. The results of the sequential optimization required at least five molecules to create a molecular representation consistent with experimental data; while a starting population of 50 molecules was required for a "high-quality" molecular representation. This is in sharp contrast to the 10,000 required for the scheme developed by Neurock *et al*. The calculated molecular weights were in good agreement with values obtained using VPO (Fig. 2-6).

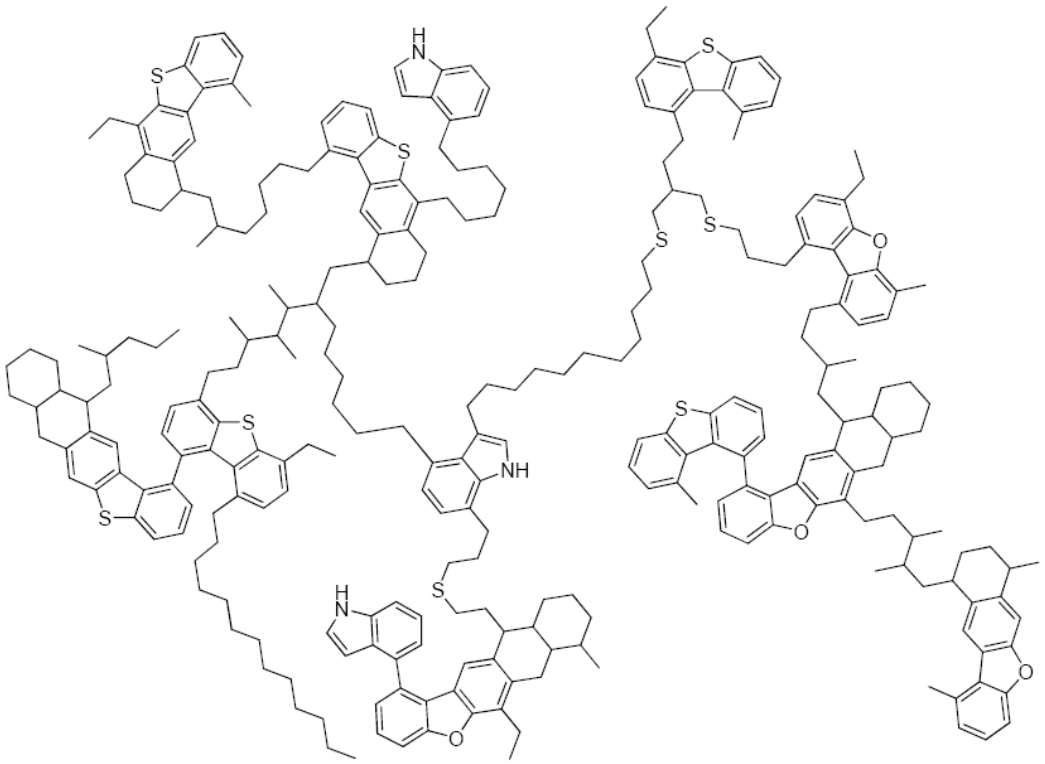


Figure 2-6: A model molecular representations of asphaltenes from Sheremata *et al.* [46]. The elemental composition and molecular weight are given as $C_{318}H_{395}N_6O_6S_8V$, and 4133 g/mol respectively.

The molecules generated were composed of branched but not cross-linked aromatic and aliphatic bridges due to the absence of analytical data supporting the existence of the latter [43].

In a recent publication, Boek *et al.* further extended the Sheremata model to create a generic asphaltene model by allowing for both three-dimensional and two-dimensional asphaltene structures [44]. Their model allowed for the generation of both archipelago and pericondensed molecular representations of asphaltenes. Aromatic sheets and aliphatic chains were used as the building blocks for generating the asphaltene molecule. The number of benzene rings was used to denote the size attribute for the aromatic sheets, while the chain-length (not necessarily the number of carbons) represented the size attribute of the aliphatic building blocks [44]. The aromatic building blocks used in the study are presented in Fig. 2-7.

The building blocks were randomly sampled and linked together using a connection algorithm which allowed for both pericondensed and archipelago structures. A nonlinear optimization procedure was used to select a subset of molecules that provided the best agreement with experimental ^{1}H, ^{13}C, elemental analysis, and molecular weight (MW) data. They tested MW values of 750 g/mol and 4190 g/mol representing pericondensed and archipelago structures respectively. The results reported showed that the peri-condensed representation was more consistent with experimental MW data. The authors contend that Sheremata's model was flawed because it had been constrained to generate solely archipelago structures; while the flexibility of their model ensures that the results obtained were more consistent with experimental data.

Consequently, the use of advanced stochastic algorithms such as Monte Carlo simulation coupled with experimental NMR (^{1}H and ^{13}C), elemental analysis, and MW data to generate representative molecules of complex heavy petroleum

feedstocks has not resolved the ambiguity surrounding the nominal average molecular structure of asphaltenes.

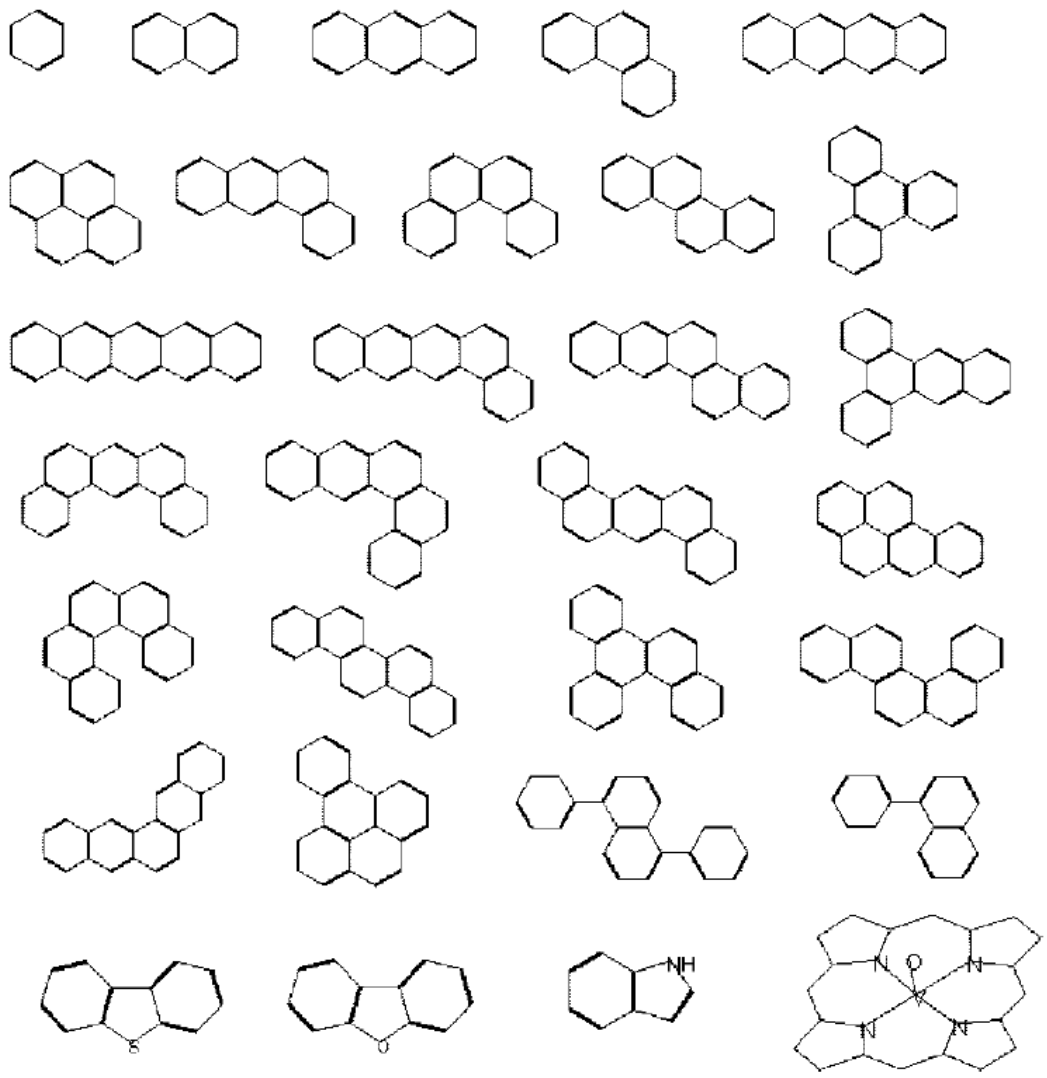


Figure 2-7: Schematic representation of aromatic groups used in the molecular representations of Boek *et al.* [44].

2.7 Vibrational Spectroscopy

Infrared (IR) and Raman spectra are both formed as a result of the absorption of electromagnetic radiation at frequencies that correspond to the vibration of specific sets of chemical bonds within a molecule [68]. The requirement for infrared activity is a net change in dipole moment during vibration of the molecules, while a net change in bond polarizability must be observed for Raman activity [68]. For molecules with a center of symmetry, fundamental frequencies that appear in the infrared spectrum are absent from the Raman, and vice-versa; whereas for molecules without a center of symmetry, most vibration frequencies are observed in both infrared and Raman spectra [13, 69].

IR and Raman spectroscopy enable the identification of aromatic, naphthenic, and paraffinic [70] structures because each of these structures vibrates at specified frequencies. Notably, far-infrared (FIR) vibrations occurring at very low frequencies (or wavenumbers) are very important because they represent skeletal vibrations [69], thus they are molecule-dependent and provide unambiguous fingerprints of specific molecular structures [71, 72]. The complementary nature of both provide an almost complete overview of the vibrational behaviour of molecules [69]. Infrared spectroscopy has been applied to the accurate identification of polycyclic aromatic hydrocarbons[71, 73], which are abundant in asphaltenes.

Yen *et al*. applied infrared spectroscopy to the elucidation of the structure of petroleum asphaltenes [74]. Their work focused primarily on developing a better understanding of the structure of the saturated components of the non-hydrocarbon asphaltic fraction of crude oils. By analyzing the frequency bands corresponding to various C-H vibrations, they were able to estimate the number of carbons existing as naphthenic, paraffinic, and terminal methyl groups. Additional structural details including polynuclear aromatic hydrocarbons were also

estimated using a correlation between structure and spectra. The results of their studied showed that condensed naphthenic rings are present in asphaltenes, while the aromatic clusters appear to be pericondensed.

Wilt *et al*. [75] used a combination of partial least squares (PLS) and Fourier transform infrared spectroscopy (FT-IR) to quantitatively determine the asphaltene content of petroleum crude oils faster than conventional solvent-based methods [75]. Using a training data set of 42 and a predictive data set of 8 crude oil samples they developed a model that resulted in a coefficient of determination (R^2) of 0.95 and 0.96 respectively.

Blanco *et al*. reported on the use of near infrared (NIR) spectroscopy to predict the physico-chemical properties of 66 various bitumen samples in 2001 [76]. They applied PLS regression to the NIR spectrum of these samples to determine their viscosity, density, softening point, and chemical composition. The values for density reported showed a 0.2% relative standard deviation from the ASTM D-70 values [76].

More recently, Hannisdal *et al.* [77] and Aske *et al.* [78] used a combination of multivariate analysis and vibrational spectroscopy to perform group-type analysis of crude oils. The latter performed their analysis using eighteen light crude oil and condensate samples, while twenty heavy crude oil samples were analyzed by the former. Both concluded however, that a combination of PLS, principal component analysis (PCA), IR, and NIR offered a faster alternative to determining the saturates, aromatics, resins, and asphaltenes (SARA) contents of heavy and light crude oils than conventional fractionation methods. As evidenced from the above, the use of vibrational spectroscopic techniques in the structural analysis of petroleum fractions including heavy oils and asphaltenes is a very active area of research.

The determination of vibrational frequencies has been greatly enhanced by quantum-mechanical computational methods such as DFT. The theoretically predicted frequencies can serve as fingerprints for the identification of experimentally observed reactive intermediaries, and the derivation of thermochemical information [13, 79]

In this work, the DFT technique as implemented in the commercial molecular simulation package Gaussian 03 [80] is used to compute vibrational frequencies (IR and Raman inclusive). This software package can compute the vibration spectra of molecules both in the ground and excited states using various DFT functionals and basis sets[81]. A combination of the three parameter DFT exchange functional of Becke-Lee-Yang-Parr [B3-LYP] and the 6-311G basis set was selected because this combination yields vibrational frequencies that are in close agreement with experimental values [79, 82-85].

2.8 Density Functional Theory (DFT)

Density Functional Theory (DFT) is a primarily theory of the electronic structure of atoms, molecules and solids in their ground states; and represents one of the most successful approaches used in the determination of the electronic structure of atoms, molecules and solids [86]. The theory depends on the 3-dimensional electron density distribution rather than the 3N-dimensional (where N is the number of electrons) electron wave function [87, 88]. It reduces the complex many-body wave function, which is the solution of the Schrodinger equation, to an effective one-body system represented by the electron density [87]. Several excellent molecular quantum chemistry texts have been written that provide a clear exposition on the theoretical development of this theory. Some of these texts include *Essentials of Computational Chemistry: Theories and Models* by Christopher J. Cramer [89], *Modern Chemistry: Introduction to Advanced Electronic Structure Theory*, by Szabo and Ostlund [88].

However, for the purposes of this work, a brief overview of the key concepts in the development of DFT is presented.

2.8.1 Quantum Mechanical Many-Electron Problem

The term "many-electron problem" describes the intractable nature of finding an exact mathematical solution for the Schrödinger equation for a system of interacting electrons.

The complete non-relativistic Schrödinger equation for a system of *n* electrons and *N* nuclei is an eigenvalue equation, the solution of which yields information about all the properties of the electrons of an atomic or molecular system, is given by:

$$\hat{H}\Psi(\mathbf{R},\mathbf{r}) = E\Psi(\mathbf{R},\mathbf{r}) \tag{2.1}$$

where E is the energy of a given eigenstate, and $\hat{\mathrm{H}}$, the Hamiltonian operator, representing the total energy of the system is

$$\hat{H} = -\frac{h^2}{2}\sum_{I}^{N}\frac{\nabla_I^2}{m_I} - \frac{h^2}{2m_e}\sum_{I}^{N}\nabla_i^2 - \sum_{i}^{n}\sum_{I}^{N}\frac{Z_I e'^2}{r_{Ii}} + \sum_{I>J}^{N}\frac{Z_I Z_J e'^2}{r_{IJ}} + \sum_{i<j}^{N}\frac{e'^2}{r_{ij}}, \tag{2.2}$$

and for simplicity

$$e'^2 = \frac{e^2}{4\pi\varepsilon_0} \tag{2.3}$$

The lower and upper case indices (2.2) represent the electrons and nuclei respectively, h is Planck's constant, m_e is the electron mass, m_I is the mass of the nucleus, and r is the distance between the objects specified by the subscript [90]. The first and second terms represent the kinetic energy of the nuclei and electrons respectively; the third term represents electron-nucleus interaction energy; the

fourth and fifth terms represent the energies due nucleus-nucleus and electron-electron interactions respectively. The solution of equation (2.1) , $\psi(\mathbf{R},\mathbf{r})$ is the many-body wave function for all *n* electrons (**r** represents electron positions) and *N* nuclei (**R** represents nuclear positions) in the system; and is a function of the position vector (**R, r**) where (**R, r**) = ($\mathbf{R}_1$, $\mathbf{R}_2$,…,$\mathbf{R}_N$, $\mathbf{r}_1$,$\mathbf{r}_2$,…,$\mathbf{r}_n$) .

The Born-Oppenheimer approximation allows for the separation of the total wavefunction into an electronic and nuclear component, thus simplifying the solution of equation (2.1).

$$\Psi(\mathbf{R},\mathbf{r}) = \Phi(\mathbf{R})\psi(\mathbf{r}) \tag{2.4}$$

This approximation is possible because the nucleus is several orders of magnitude heavier than the electrons, and can be treated as stationary relative to the motion of the electrons. Hence, the solution of (2.1) now depends on the motion of the electrons in the system in general described by the many-electron wavefunction $\psi(\mathbf{r})$, and the difficult electron-electron interaction term in particular; hence, the "many-electron problem".

The solution of the difficult electron-electron interaction term is the subject of several electronic structure theories. Attempts to find approximate solutions to this term and by extension the many-body Schrödinger equation will be discussed in the following sections.

2.8.2 The Hartree-Fock Approximation

The Hartree and Hartree-Fock theories were developed in the 1930's [91] to solve the electronic Schrödinger equation obtained after the application of the Born-Oppenheimer approximation described in the previous section. Since the electronic wave function, $\psi(\mathbf{r})$ depends on the coordinates of all the electrons in

the system, Hartree proposed that the wavefunction can be separated into a product of one-electron wavefunctions [90]. In order to solve the Schrödinger equation, Hartree replaced the electron-electron term with an effective potential; which represents the repulsion experienced by one electron as a result of the presence of other electrons [90]. With this assumption, the single-electron equation of an atom in Hartree's theory is

$$\psi(\mathbf{r}_1, \mathbf{r}_2 ... \mathbf{r}_n) = \phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2)...\phi_n(\mathbf{r}_n) \tag{2.5}$$

$$\left(-\frac{1}{2}\nabla_i^2 - \sum_I^N \frac{Z_I}{r_{Ii}} + V_i^{eff}\right)\phi_i = E_i\phi_i \tag{2.6}$$

where ϕ_i represents the separable single-electron wavefunctions, and V_i^{eff} is the effective potential.

Fock observed that the Hartree Product (2.5) did not satisfy the antisymmetric principle (Pauli Exclusion Principle); which states that a wavefunction describing an electron must be antisymmetric (result in a sign change) with respect to the interchange of any set of spin-state coordinates [92]. To address this shortcoming, he suggested the use of a Slater determinant to represent the single-electron wavefunctions, and also retained the effective potential from Hartree's equation,

$$\psi(\mathbf{r}_1, \mathbf{r}_2 ... \mathbf{r}_n) = \frac{1}{\sqrt{n!}}\begin{vmatrix} \phi_1(e_1) & \phi_2(e_1) & \dots & \phi_n(e_1) \\ \phi_1(e_2) & \phi_2(e_2) & \dots & \phi_n(e_2) \\ \vdots & \vdots & \cdots & \vdots \\ \phi_1(e_n) & \phi_2(e_n) & \dots & \phi_n(e_n) \end{vmatrix} = \left|\phi_1, \phi_2 \dots \phi_n\right| \tag{2.7}$$

The equations arising from the Hartree-Fock theory need to be solved self-consistently since the effective potential in both equations are dependent on the single-particle wavefunctions ϕ_i .The variational principle states that the best

product many-electron wavefunction is the one with the minimum expectation value of the energy; this condition can be used to derive equations for the best one-electron wavefunctions [90].

The Hartree-Fock theory is also referred to as an *independent particle model* or *mean field theory*, since the assumption of an antisymmetrized product (Slater determinant) description of the electrons implies that each electron is independent of the others but experiences Coulomb repulsions due to their presence [92]; however, it still gives reliable results for many-body problems [93].

2.8.3 The Kohn-Sham Equations

In 1964, Hohenberg and Kohn (HK) [94] published a paper that laid the foundations for the development of DFT. The paper dealt primarily with the development of a theorem for the ground state of an interacting electron gas in an external potential. According to HK, the specification of the ground state electron density $n(\mathbf{r})$ determines the external potential $v(\mathrm{r})$ *uniquely* [95] . Since $n(\mathrm{r})$ also determines the total number of electrons, N, by integration, it determines the full Hamiltonian H, and consequently all properties determined by H; including the full many-body wavefunction, Ψ [95].

The second HK theorem states that there exists a functional, $\boldsymbol{E}[n(\mathbf{r})]$, of the electron density, $n(\mathbf{r})$, such that the exact ground state energy is the global minimum of this functional. The minimal principle for a given potential $v(\mathrm{r})$ can be expressed as [95],

$$E_{v(r)}[n(r)] \geq E_{v(r)}[n_0(r)] \equiv E_0 \tag{2.8}$$

where $n_0(r)$ and E_0 are the density and energy of the ground state. This equality holds only if $n_0(r) = n(r)$

The energy functional of $n(\mathbf{r})$ is defined as

$$E_{v(r)}[n(r)] \equiv \int v(r)n(r)dr + F[n(r)] \tag{2.9}$$

where

$$F[n(r)] \equiv T_s[n(r)] + \frac{1}{2}\int \frac{n(r)n(r')}{|r-r'|} drdr' + E_{XC}[n(r)] \tag{2.10}$$

$T_s[n(r)]$ is the ground state kinetic energy of a set of *noninteracting* electrons, the next term is their interaction energy, and $E_{XC}[n(r)]$ is the exchange-correlation energy.

Utilizing the HK theorems, the energy functional, $E_{v(r)}[n(r)]$ is minimized subject to the condition $\int n(r)dr = N$ (which implies that the total number of electrons in the system is constant), to obtain the orbitals that give rise to the ground state energy. This leads to a set of equations called the Kohn-Sham (KS) [96] equations

$$(-\frac{1}{2}\nabla^2 + v(r) + \int \frac{n(r')}{|r-r'|} dr' + v_{XC}(r) - \varepsilon_j)\varphi_j = 0 \tag{2.11}$$

$$n(r) = \sum_{j=1}^{N} |\phi_j(r)|^2 \tag{2.12}$$

$$v_{XC}(r) = \delta E_{XC}[n(r)] / \delta n(r) \tag{2.13}$$

These equations must be solved self-consistently for ε_j and φ_j with an appropriate approximation for the exchange correlation energy $E_{XC}[n(r)]$; thus representing the many-particle system in terms of single-particle orbitals. The KS orbitals are appropriate for noninteracting electrons, and by introducing the modified potential the exact electron density for a system with full interaction is

obtained [97].These equations facilitate the reduction of the many-body electron problem to a set of single-particle orbitals, and form the basis for the DFT [98]. Much of the research in DFT has involved the development of approximations for the exchange correlation energy term. One of the most successful approximations is the Becke, Lee, Yang, and Parr functional also referred to as the BLYP functional [91].

A density functional calculation involves the selection of an appropriate functional with a basis set. When a set of single-electron wavefunctions from individual atoms are used to construct more complex molecular systems by using a linear superposition of the atomic orbitals centered on the individual atoms, the set of *N* equations generated is called a basis set [89]. The combination B3LYP/6-311G uses a B3LYP functional with a 6-311G basis set.

2.9 Structural Units in Petroleum

The chemical complexity and diversity of petroleum fractions including heavy oil, bitumen, vacuum residue and asphaltenes makes it impossible to synthesize model compounds that completely describe their physical, chemical and thermodynamic properties. However, several studies have identified the chemical structures (Fig. 2-8) that are present in petroleum [3, 23, 39, 57, 59, 61, 99]. The theoretical vibrational spectra of some of these structures have been studied in this work. In addition, varying lengths of aliphatic chains were also used to model the aliphatic bridges linking aromatic rings that are characteristic of the observed structure of asphaltenes [43, 57].

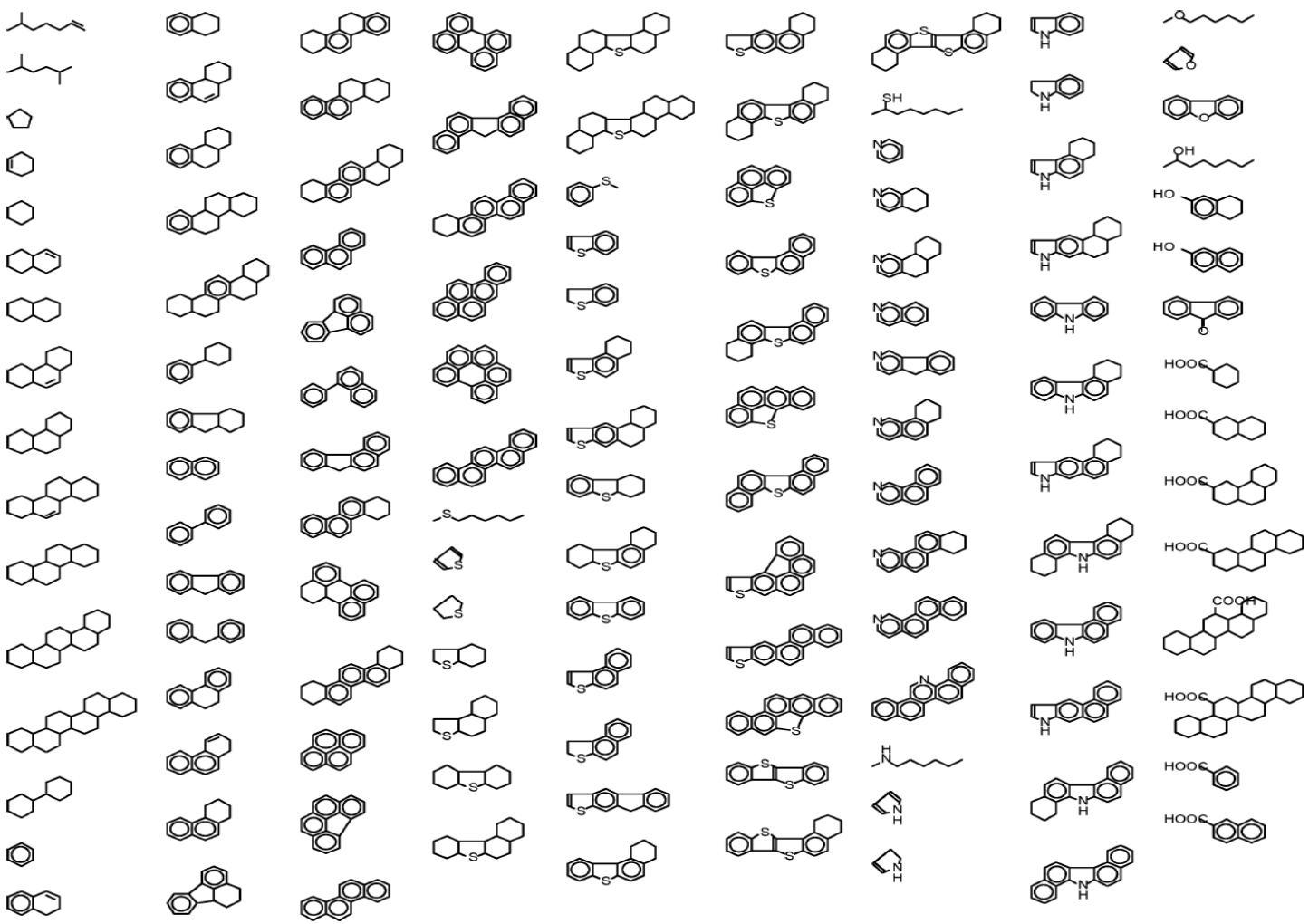


Figure 2-8: Depiction of the known core structures identified in petroleum[61]

2.10 Summary

The foregoing review highlights the substantial efforts made to elucidate the molecular structure of ill-defined hydrocarbons. Definition of the asphaltene fraction in particular provides a key illustrative example because for heavy oil, bitumen, and vacuum residue asphaltenes are a principal constituent. Hence to understand these ill-defined petroleum fractions, an understanding of the physico-chemical and thermodynamic properties of asphaltenes is absolutely essential.

Molecular structure is an important input for understanding these properties. So far both peri-condensed and archipelago molecular structures have been proposed. However, it is interesting to note that both sets of structures have been proposed using experimental data obtained using exactly the same set of analytical characterization techniques namely, ^{1}H and ^{13}C NMR, mass spectroscopy, and elemental composition; as well as the same statistical analysis techniques[43, 44].

This ambiguity prompts the hypothesis that ^{1}H and ^{13}C NMR provide information on too small a length scale to yield unambiguous molecule scale structures. A second hypothesis explored in this work is that techniques providing molecular information on larger, but still sub-molecular length scales are required to discriminate between the two molecular structural prototypes in particular and for ill-defined hydrocarbon species in general. In this work a combination of infra-red and Raman spectroscopy is evaluated as a potential basis for unambiguous molecular structure definition, in addition to or in lieu of ^{1}H and ^{13}C NMR, mass spectroscopy.

3 Computational Approach

3.1 Introduction

The development of computational science has enabled the detailed analysis of complex physical and chemical systems that are either too expensive to observe empirically or intractable to solve analytically. Computational techniques have benefited immensely from the growth in the computational speed and accuracy of modern computers; this has provided experimentalists the means to create more accurate models of real-life systems, and simulate various possible scenarios before conducting actual laboratory experiments.

As mentioned in the previous chapter, the many-electron problem is one of the most intractable problems in quantum physics. The distribution and motion of electrons about the nuclei are governed by the Schrödinger equation, but finding a solution to this equation is very difficult. This difficulty is due to the presence of the electron-electron interaction term, which makes it impossible to separate the many-electron system into one consisting of many single-electron systems. Any attempt to probe the structure of atomic and molecular systems thus requires a solution of the electron-electron interaction term of the Schrödinger equation.

A major contribution to the approximate solution of the Schrödinger equation involved the separation of the nuclei and electronic motion, also known as the Born-Oppenheimer approximation. This enabled the electronic component to be solved with the nuclear positions as parameter; thus the major computational effort is in solving the electronic Schrödinger equation for a given set of nuclear

coordinates. The Hartree-Fock (HF) theory approximated the electron-electron interaction by an average potential, thus neglecting the correlation between electrons [100]. However, the theory laid the foundation for the development of more accurate electronic structure methods. It represented the many-electron wavefunction solution of the Schrödinger equation as a product of one-electron orbitals; the Slater determinant was used to represent these orbitals in order to account for antisymmetry of electron spins.

The Density Functional Theory (DFT) can be considered as an improvement on the HF theory, where the many-body effect of electron correlation is modeled using a functional of the electron density; it is comparable to HF but provides significantly better results. However, while the HF model can be improved to converge towards the exact solution of the Schrödinger equation by adding additional determinants, there is no systematic approach to improving the DFT results [100].

The computational approach applied in this work involved four key steps,

1) Selection of an appropriate exchange-correlation functional
2) Choosing the right basis set
3) Frequency calculation, and
4) Linear Regression Analysis.

The following sections discuss each step in more detail.

3.2 Electronic Structure Methods

Electronic structure methods are mathematical approximations to the solutions of the Schrödinger equation. There are two classes of electronic structure methods: *Ab initio* and Semi-empirical methods.

Semi-empirical methods use parameters derived from experimental data to simplify the solution of the Schrödinger equation; these parameters are dependent on the type of chemical system under investigation. *Ab initio* methods unlike the semi-empirical methods do not depend on experimental parameters; their computations are based entirely on the laws of quantum mechanics and solutions to the Schrödinger equation are obtained using a series of rigorous mathematical approximations [81].

Semi-empirical methods are relatively inexpensive computationally and provide reasonable descriptions of molecular systems and fairly accurate quantitative predictions of molecular properties such as energies and structures for systems with good parameters. However, *ab initio* methods provide high quality predictions for a broad range of molecular systems, and are not limited to any specific system classes [81]. The development of large supercomputing capacity has facilitated the application of *ab initio* methods to large, complex molecular systems.

The Hartree-Fock theory is the least expensive *ab initio* method available, but the Density Functional methods provide much better results with the same amount of computational resources [81]. While DFT methods account for the correlation effects arising from electronic motion, the HF calculations treat this effect as an average interaction [100]; this approximation makes the HF results to be less accurate than the DFT for certain types of systems with very strong electron correlation [81].

3.3 Exchange and Correlation Functionals

The fundamental difference between DFT and other electronic structure methods is the use of the electron density instead of the electronic wavefunction to obtain all the ground state properties of a quantum-mechanical system as proposed by Hohenberg and Kohn[94]. The success of modern DFT methods is based on the Kohn-Sham[96] equations which provide for the calculation of the electron kinetic energy from an auxiliary set of orbitals used for representing the electron density[100].

The exchange-correlation energy arises because the motion of electrons through the electron density is not random but avoid one another ;this results in the creation of electron-holes around each electron (each electron is surrounded by its exchange-correlation hole density) [98]. It consists of three contributions namely, the potential energy of the exchange, the potential energy of the correlation, and the smaller kinetic energy of the correlation [98].

As mentioned in Chapter 2, DFT can only be practically useful by generating good approximations for the exchange-correlation energy term $E_{XC}[n(r)]$; which is a relatively small component of the total energy of a typical system [98, 100].

The exchange-correlation energy can be separated into exchange and correlation functional components [101]

$$E_{XC}[n(r)] = E_X[n(r)] + E_C[n(r)] \tag{3.1}$$

The two functionals can depend only on the electron density, *n(r)* hence local density functionals and/or on the gradient-corrected functionals that depends on the both density, *n(r)* and its gradient $\nabla n(r)$ [101].

3.3.1 Local Density Approximations (LDA)

The simplest exchange-correlation energy model is the local density approximation (LDA) also called Local Spin-Density Approximations (LSDA) [98] where the electron density is assumed to be varying slowly enough for the exchange-correlation energy to be calculated using equations derived for a uniform electron density [100].

$$E_{XC}^{LDA}[n(r)] \equiv \int \varepsilon_{XC}(n(r))n(r)dr \tag{3.2}$$

where $\varepsilon_{XC}(n(r))$ is the exchange-correlation energy per particle of a uniform interacting electron gas of density *n(r)*. The KS orbitals in the LDA approximation are usually very close to the HF orbitals [95]; however the LDA underestimates the exchange energies and of atomic and molecular systems by roughly 10% [102].

3.3.2 Generalized Gradient Approximations GGA)

A significant improvement in accuracy can be obtained by making the exchange-correlation functional dependent also on the first derivative of the electron density [98, 100]; this level of approximation is called the generalized gradient approximations (GGA)

$$E_{XC}^{GGA} = \int f(n(r),|\nabla n(r)|)dr \tag{3.3}$$

where $f(n(r),|\nabla n(r)|)$ is a suitably chosen function of its two variables [95].

The GGA exchange functional developed by Becke [103] is regarded as the first to be widely used, and is denoted by "B" [89]; while the Lee, Yang, and Parr (LYP) [104] GGA functional has been widely used for the correlation energy [100]. A complete specification of the exchange and correlation functionals is achieved by concatenating the two acronyms in that order; such that a BLYP calculation combines Becke's GGA exchange with the GGA correlation functional of Lee, Yang, and Parr [89].

3.3.3 Hybrid Functionals

Becke [103, 105] developed hybrid functionals which arise from the combination of Hartree-Fock, local, non-local, and gradient-corrected exchange functionals [101] to define the exchange-correlation energy. Conceptually, the exchange-correlation is defined as a mixture of Hartree-Fock and DFT exchange along with DFT correlation [81], i.e.

$$E_{XC} = c^{HF} E_X^{HF} + c^{DFT} E_{XC}^{DFT} \tag{3.4}$$

where the *c's* are constants to be determined.

For example, the exchange-correlation functional proposed by Becke [103, 106] is of the form

$$E_{XC} = E_{XC}^{LSDA} + a_0(E_X^{HF} - E_X^{LSDA}) + a_X \Delta E_X^{B88} + a_C \Delta E_C^{PW91} \tag{3.5}$$

Here ΔE_X^{B88} is Becke's 1988 gradient correction (to the LSDA) [103] for exchange functional, and ΔE_C^{PW91} is the 1991 Perdew-Wang [103, 107] gradient

correction to the correlational functional, E_X^{HF} is the Hartree-Fock (*exact*) exchange [103]. Becke suggested coefficients $a_0 = 0.2$, $a_X = 0.72$, and $a_C = 0.81$ based on fitting to heats of formation of small molecules[106]. This is known as the Becke three-parameter functional. The Becke3LYP or B3LYP functional uses LYP instead of PW91 for the gradient-corrected correlation functional [103, 106]. The B3LYP implemented in Gaussian [80, 106] has the form,

$$E_{XC}^{B3LYP} = (1-a_0)E_X^{LSDA} + a_0 E_X^{HF} + a_X \Delta E_X^{B88} + a_C E_C^{LYP} + (1-a_C)E_C^{VWN} \qquad (3.6)$$

where all the constants have the same meaning as in equation (3.5) but VWN local correlation functional has been included because the LYP functional does not have easily separable local component [106].

The B3LYP functional and hybrid functionals in general have been shown to produce vibrational force fields, frequencies, spectra, and thermochemical properties in excellent agreement with experiments; and their level of accuracy is far superior to those of LSDA and BLYP functionals [106].

3.4 Basis Set

A basis set is an approximation that facilitates the solution of *ab initio* calculations [100]; it is a mathematical representation of the orbitals of a molecular system [81]. In quantum mechanics, electrons have a finite probability of existing anywhere in space; hence a complete basis set is one that can span the entire space. This would require an infinite number of functions, which is computationally impossible [81, 100].

Two major factors influence basis set accuracy (1) the size of the basis set [81], and (2) the type of basis set functions used [100]. The smaller the basis set used,

the poorer the representation of the molecular orbitals [100], hence, larger basis sets more accurately approximate the orbitals by imposing fewer restrictions on the locations of the electrons in space [81]. Furthermore, the better a single basis function is able to reproduce or represent the orbital, the fewer the number of basis functions required to generate a more accurate approximation [100]

There are two types of basis functions commonly used in electronic structure calculations: *Slater Type Orbitals* (STO) and *Gaussian Type Orbitals* (GTO) [100]. The functional forms of STO's and GTO's are given in equations (3.7) and (3.8) respectively

$$\chi_{\zeta,n,l.m}(r,\theta,\varphi) = NY_{l,m}(\theta,\varphi)r^{n-1}e^{-\zeta r} \tag{3.7}$$

$$\chi_{\zeta,n,l.m}(r,\theta,\varphi) = NY_{l,m}(\theta,\varphi)r^{2n-2-l}e^{-\zeta r^2} \tag{3.8}$$

Here N is normalization constant, and $Y_{l,m}$ are spherical harmonic functions. As a rule of thumb, three GTO's are usually required to achieve the same level of accuracy as one STO [100]. A linear combination of Gaussian functions that represents one STO is referred to as a *contracted Gaussian function*, and the individual Gaussian functions in the combination are called *primitive Gaussians*[91]; and a basis set consisting of a single Gaussian function is termed *uncontracted* [81]. However, even though a larger number of GTO basis functions are required, their computational efficiency makes them the preferred basis sets for electronic structure calculations [81, 100].

For a more comprehensive discussion on the this subject, please refer to *Introduction to Computational Chemistry* by Frank Jensen [100], and *Quantum Chemistry* by Donald A. McQuarrie [91]

3.4.1 Split Valence Basis Sets

Basis sets assign a group of basis functions to each atom within a molecule to approximate its orbitals [81]; and the type and number of functions used depends on the level of accuracy required. *Minimal basis sets* contain the minimum number of basis functions required for each atom. An example of a minimal basis set is the STO-3G basis set (which uses a contraction of three Gaussian functions to represent one Slater type orbital, thus the "3G" in its name). A basis set can be made larger by increasing the number of basis functions per atom. *Double Zeta* (DZ) *basis sets* generate the molecular orbitals by doubling the sizes of the basis functions for each atomic orbital. A variation of the double zeta basis set is the *split valence basis set;* this describes the core electrons by a single orbital and the valence shell electrons by a sum of orbitals [91, 100]. *Split valence basis sets* allow orbitals to change size , and *triple split valence* basis sets, like 6-311G, use three sizes of contracted Gaussian functions for each orbital-type [81].

There are *Quadruple Zeta* (QZ), and *Quintuple* or *Pentuple Zeta* (PZ or 5Z) basis sets; and each represents the number of basis functions used for each orbital-type [100]. Polarized basis sets allow orbitals to change shape by adding orbitals with angular momentum functions, while basis sets with diffuse functions allow orbitals to occupy a large region of space thus accounting for systems in their excited states, and molecules with lone pairs [81].

Several publications [13, 108, 109] have applied the 6-311G basis set in combination with the B3LYP exchange-correlation functional to obtain vibrational frequency data in very good agreement with experimental values. Hence, the vibrational frequency data used in this work have been generated using this same combination

3.5 Frequency Calculation

After selecting the appropriate exchange-correlation functional and basis set, the next step is generating the vibrational frequency data. According to the Born-Oppenheimer approximation, electronic and nuclear motions can be decoupled in the Schrodinger equation. The resulting electronic Schrodinger equation is then solved for a large number of nuclear geometries to give the *potential energy surface* (PES) [89, 100]. The rotational and vibrational motions of a given set of nuclei, *N* on the PES is described by *3N* nuclear coordinates; three coordinates each describe the overall translational and vibrational motions with respect to three axes [100].

The PES is a hypersurface defined by the potential energy of a collection of atoms over all possible atomic arrangements, and it provides complete information about all the possible chemical structures and all isomerization pathways for a given collection of atoms [89]. Furthermore, the PES describes how the energy of a molecular system varies with small changes in its structure [81].

Molecular frequencies depend on the second derivative of the energy with respect to the nuclear positions; hence frequency calculations are only valid at stationary points on the PES, and must be performed on optimized structures [81].

Geometry optimizations are performed to locate minima on the PES, thereby predicting equilibrium structures of molecular systems. The first derivative of the energy or gradient indicates the direction along the PES in which the energy experiences the most rapid decrease, and it is the negative of the forces on the PES. All successful optimizations seek to locate the *minima* or stationary points. At the minima, both the gradient and forces are essentially zero; at this point the optimization is complete or converged [81].

The frequency calculation and the optimization are both conducted at the same level of theory and basis set; otherwise the results have no validity. Gaussian [80] can compute the vibrational spectra of molecules in their ground and excited states; and analytic second derivative of the energy with respect to nuclear positions are available for the B3LYP DFT theoretical models [81] .

3.6 Linear Regression Analysis

Exploring relationships between two or more variables forms the basis of many engineering and scientific investigations. For instance in engineering thermodynamics, establishing the relationship between the phase behaviour of petroleum fluids and changes in pressure, temperature and composition is crucial to developing reliable equations of state models; which in turn improves the design of commercial process simulators and more efficient chemical process design.

Regression analysis is a statistical technique that can be used to investigate relationships between two or more variables. The technique involves approximating the relationship between a *response* variable and an *explanatory* variable or variables (also called *regressors* or *predictors*). If this relationship can be defined by a linear function, then it is a *linear* regression, otherwise it is *nonlinear*. A regression model that contains more than one explanatory variable is referred to as a *multiple regression model* [110]; it is called a *simple regression model* if only one regressor is used.

A multiple linear regression model describing the relationship between a *response* variable, Y, and k *regressor* variables, x_k can be represented by

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k + \varepsilon \tag{3.9}$$

The parameters α_j, $j = 0, 1,\ldots, n$ are called regression coefficients; and they represent the expected change in response Y per unit change in x_j when all the remaining regressors x_i ($i \neq j$) are held constant [110]. ε is a random error term.

3.6.1 Method of Least Squares

The regression coefficients in equation 3.1 can be estimated by the *method of least squares* [110]. Suppose there are $n > k$ observations, such that x_{ij} represents the ith observation of variable x_j as shown in Table 3.1

Table 3-1: Sample data for *Multiple Linear Regression* (from reference 105)

y	x_1	x_2	$\cdots$	x_k
y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Each observation ($x_{i1}, x_{i2},\ldots, x_{ik}, y_i$), satisfies the model in equation 3.1, i.e.

$$\begin{aligned} y_i &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_k x_{ik} + \varepsilon_i \\ &= \alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ij} + \varepsilon_i \qquad i = 1, 2, \ldots, n \end{aligned} \tag{3.10}$$

The least squares function is

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \alpha_0 - \sum_{j=1}^{k} \alpha_j x_{ij} \right)^2 \tag{3.11}$$

The minimization of L with respect to the *least square estimates* [110] of α_0, α_1,…, α_k, say $\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_k$ must satisfy

$$\left.\frac{\partial L}{\partial \alpha_0}\right|_{\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\alpha}_0 - \sum_{j=1}^{k}\hat{\alpha}_j x_{ij}\right) = 0 \tag{3.12}$$

and

$$\left.\frac{\partial L}{\partial \alpha_0}\right|_{\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\alpha}_0 - \sum_{j=1}^{k}\hat{\alpha}_j x_{ij}\right)x_{ij} = 0 \quad j = 1, 2, \ldots, k \tag{3.13}$$

Simplifying these two equations, we obtain the *least squares normal equations*

$$\begin{array}{ccccccc}
n\hat{\alpha}_0 + \hat{\alpha}_1\sum_{i=1}^{n}x_{i1} & + \hat{\alpha}_2\sum_{i=1}^{n}x_{i2} & + \cdots + \hat{\alpha}_k\sum_{i=1}^{n}x_{ik} & = \sum_{i=1}^{n}y_i \\
\hat{\alpha}_0\sum_{i=1}^{n}x_{i1} + \hat{\alpha}_1\sum_{i=1}^{n}x_{i1}^2 & + \hat{\alpha}_2\sum_{i=1}^{n}x_{i1}x_{i2} & + \cdots + \hat{\alpha}_k\sum_{i=1}^{n}x_{i1}x_{ik} & = \sum_{i=1}^{n}x_{i1}y_i \\
\vdots \quad\quad \vdots & \vdots & \vdots \quad\quad \vdots & \vdots \\
\hat{\alpha}_0\sum_{i=1}^{n}x_{ik} + \hat{\alpha}_1\sum_{i=1}^{n}x_{ik}x_{i1} & + \hat{\alpha}_2\sum_{i=1}^{n}x_{ik}x_{i2} & + \cdots + \hat{\alpha}_k\sum_{i=1}^{n}x_{ik}^2 & = \sum_{i=1}^{n}x_{ik}y_i
\end{array} \tag{3.14}$$

The normal equations can be solved to obtain the *least squares estimators* [110] of the regression coefficients.

3.7 Thesis Objective

The goal of this study is to investigate the possibility of using ^{1}H and ^{13}C NMR and or IR and Raman vibrational frequency data for a library of small hydrocarbon compounds in combination with least squares regression to predict the structure of individual large molecules where all spectra are computed on the

same basis. This is equivalent to an experiment based study where measurements for all molecules are performed on the same spectrometers. This is a proof of concept study. Application to ill-defined petroleum fractions is beyond the scope of this thesis.

The large compounds (Fig. 3-2) in this study comprise two variable polynuclear aromatic groups connected by variable n-alkane chains (octane to dodecane). The vibrational frequency and NMR data of these large compounds (representing the independent variables) and the corresponding data of a library of six aromatic ring and five aliphatic chain compounds (the dependent variables) were analyzed using linear regression. The aromatic rings that make-up the library are shown in Fig. 3-1. The frequency data of the large compounds and the library of molecular fragments were obtained using a DFT level of theory and a basis set, which is known to yield IR and Raman frequency values consistent with measurement [13, 108]. Hence, all the analysis and comparisons are between computed spectra only.

The key question the study attempts to answer is the feasibility of representing the molecular structure of an *unknown* hydrocarbon compound as a linear combination of *known* molecular fragments based on vibrational frequency data. Further, the regression coefficient of each molecular fragment is used as measure of its relative abundance in the compound.

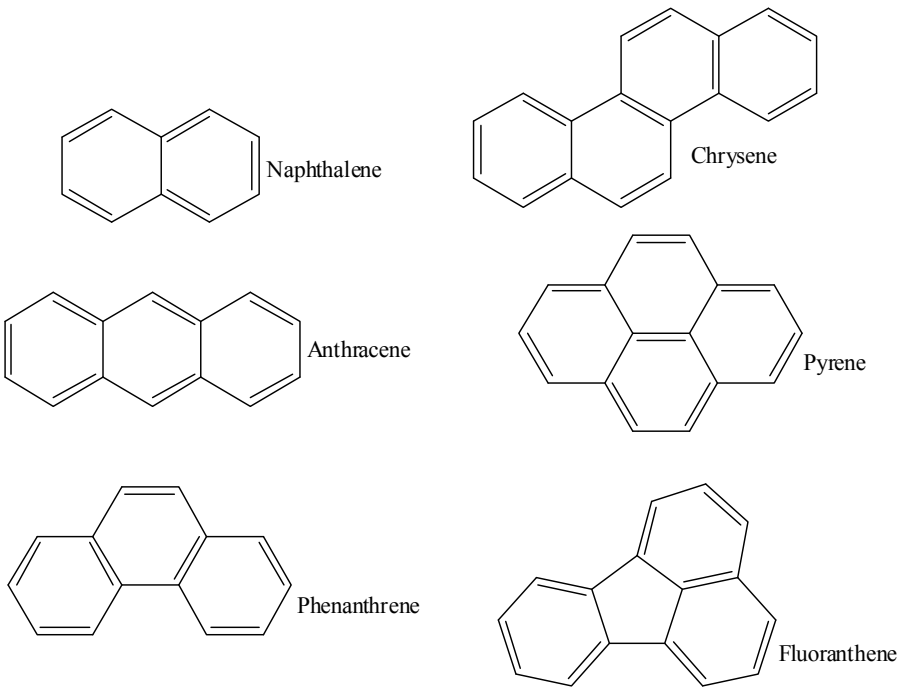


Figure 3-1: The complete library of *small* aromatic ring groups used in the study; n-alkane chains from octane to dodecane make-up the rest of the library

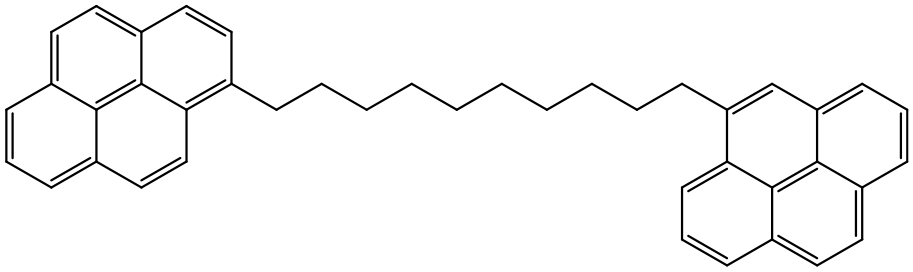


Figure 3-2: Model *large* hydrocarbon compound consisting of 2 Pyrene groups connected by an n-Decane chain

4 Molecule Identification using IR and Raman Spectroscopy

4.1 Introduction

In chapter 3, an outline of the computational methods employed in this study is provided. This chapter focuses on the specific details relating to the acquisition of the frequency data for all the hydrocarbon structural groups listed in chapter 3; as well as the model compounds generated using various combinations of these structures. In addition, the spectral optimization algorithm implemented to obtain the results is discussed.

The model compounds have been constructed using the six aromatic ring groups presented in Chapter 3, and five aliphatic chains ranging from octane to dodecane. Model compounds have been named to reflect the hydrocarbon groups they contain. For instance, a model compound composed of anthracene, decane and chrysene has been designated as *anthracene* + *decane*+ *chrysene* (Fig. 4-1). This convention has been chosen for convenience and ease of understanding and does

not constitute a deliberate disregard for the *IUPAC* (International Union of Pure and Applied Chemistry) naming conventions for chemical structures.

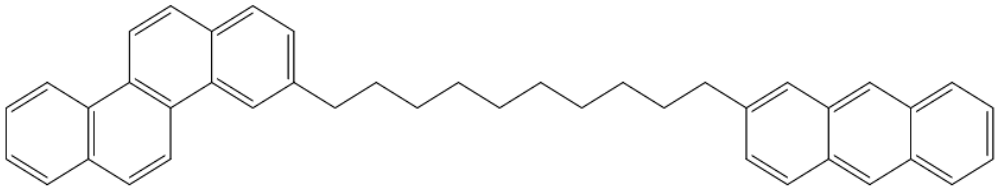


Figure 4-1: Schematic representation of a the model compound comprising Anthracene, Decane and chrysene referred to as *anthracene* + *decane* + *chrysene* following the naming convention adopted in this study.

4.1.1 Generating Frequency Data

Gaussian [80] was used to generate vibrational frequency data for this study. This software is a powerful tool that allows for the computation of the properties of molecular systems in the gas phase or in solution, as well as in their ground or excited states [81]. Since this study is exploratory in nature, the frequency data of the model compounds studied has been obtained for their ground states and in the gas phase. Furthermore, because the computation of molecular properties including vibrational frequencies is based on quantum mechanical calculations, it is possible to generate a complete frequency spectrum for molecular systems; this facilitates the application of the *spectral addition* technique employed in this study.

Frequency calculations or *jobs* in Gaussian [80] are run using an input file that contains a detailed description of the molecular structure of the compound to be investigated; and the theoretical model and basis set employed. The Gaussian [80] keyword for running a frequency calculation is ***Freq***; while the keyword for running an optimization job is ***Opt***. All the frequency calculations were performed

after an optimization procedure had been completed on the structures. A summary of the data contained in the input file is provided in Table 4-1.

Table 4-1: Input parameters for all the frequency calculations used in the study

Job Type/**Keyword**	Optimization/**Opt**			
Method	Ground State	DFT	Default Spin	B3LYP
Basis Set	6-311G			

Job Type/**Keyword**	Frequency/**Freq**			
Method	Ground State	DFT	Default Spin	B3LYP
Basis Set	6-311G			

A frequency job begins by computing the energy of the optimized input structure (which is specified in mixed Cartesian and Z-matrix coordinates); the resulting structure is the basis for computing the frequencies [81]. The computed frequencies are contained in an output file generated by Gaussian [80].

GaussView® is a software application that was used to prepare Gaussian [80] job inputs and graphically analyze the outputs. The graphical user interface design of *GaussView*® enhances the visualization of Gaussian [80] input and output files because a typical Gaussian [80] file is composed of text-only commands.

The frequency calculations were performed for 6 aromatic ring and 5 aliphatic chain compounds; as well as for an additional 20 different combinations of these compounds. The 20 combinations represented models of ill-defined hydrocarbon structures.

4.1.2 Spectral Optimization Scheme

The spectral data generated from Gaussian [80] was exported via *GaussView*® to *Microsoft Excel*® (Excel) for analysis. Using Excel's Solver (Solver) add-in, the spectral data of the combined aromatic ring and aliphatic chain model compounds (independent variable) and the library of pure aromatic ring and aliphatic chain compounds (dependent variables) were analyzed by linear regression.

The model developed to find the relationship between these two data sets consisted of three parts namely the objective function, linear regression coefficients, and model constraints. The objective function was defined as the sum of the squared difference between the spectral data of the combined aromatic ring and aliphatic chain compounds (*CAA)* and the pure aromatic ring and aliphatic chain compounds (*AA*'s); and the goal of the optimization was to minimize this difference subject to the model constraints by adjusting the linear regression coefficients. The regression coefficients were constrained to be positive values; and they provided a qualitative measure of the relative abundance of each pure *AA* in the *CAA* being investigated.

The process flow diagram in Fig. 4-2 is a depiction of the optimization algorithm employed in this study. The frequency data ranged from 0-3500 cm^{-1} with intervals of 10cm^{-1}; each *AA* in the library as well as each of the *CAA*'s investigated had infrared intensities and Raman activities for each frequency data point. The product of the linear regression coefficient (coefficient) and the infrared intensity (Raman activity) for each *AA* was summed for each frequency data point. This sum represented the *predicted* infrared intensity (Raman activity) for each frequency; while the infrared intensity (Raman activity) of the *CAA* represented the *measured* values.

The squared difference between the predicted and measured infrared intensity (Raman activity) for each frequency was summed to obtain the *objective function.*

Using Solver, this difference was minimized to find the coefficients (the final values of the regression coefficients were independent of the initial values chosen, so zero was used as the initial value).

The task of analyzing the results was greatly facilitated by the *apriori* knowledge of the structural components of the model compounds. A close observation of the results showed that most of the structural components which were absent from the model compounds had relatively low coefficients compared to those present. Hence, to eliminate spurious (or reduce the ambiguity of) results during the analysis, a threshold value of 0.30 was chosen for the coefficients.

4.1.3 Rationale for using Linear Regression Coefficients

The optimization algorithm outlined in the preceding section was performed on six model compounds consisting of two identical aromatic rings connected by n-decane; using a library consisting entirely of aromatic ring and aliphatic chain compounds present in the model compounds to test the feasibility of using linear regression coefficients to qualitatively predict the relative abundance of the various molecular sub-units in the model compounds. For instance, while analyzing *pyrene* + *decane* + *pyrene* (Fig. 4-3), the library would contain only pyrene and decane.

Since the ratio of the aromatic rings to n-decane in the model compounds is qualitatively 2:1, the coefficients should at the minimum reflect this relationship for it to be feasible. Hence, the first step in the analysis was to test this hypothesis for both infrared and Raman frequency data. The results obtained for all six cases investigated were quite satisfactory (qualitatively), and thus provided the motivation to extend the technique to a larger library of molecular sub-units; as well as various aromatic ring and aliphatic chain combinations.

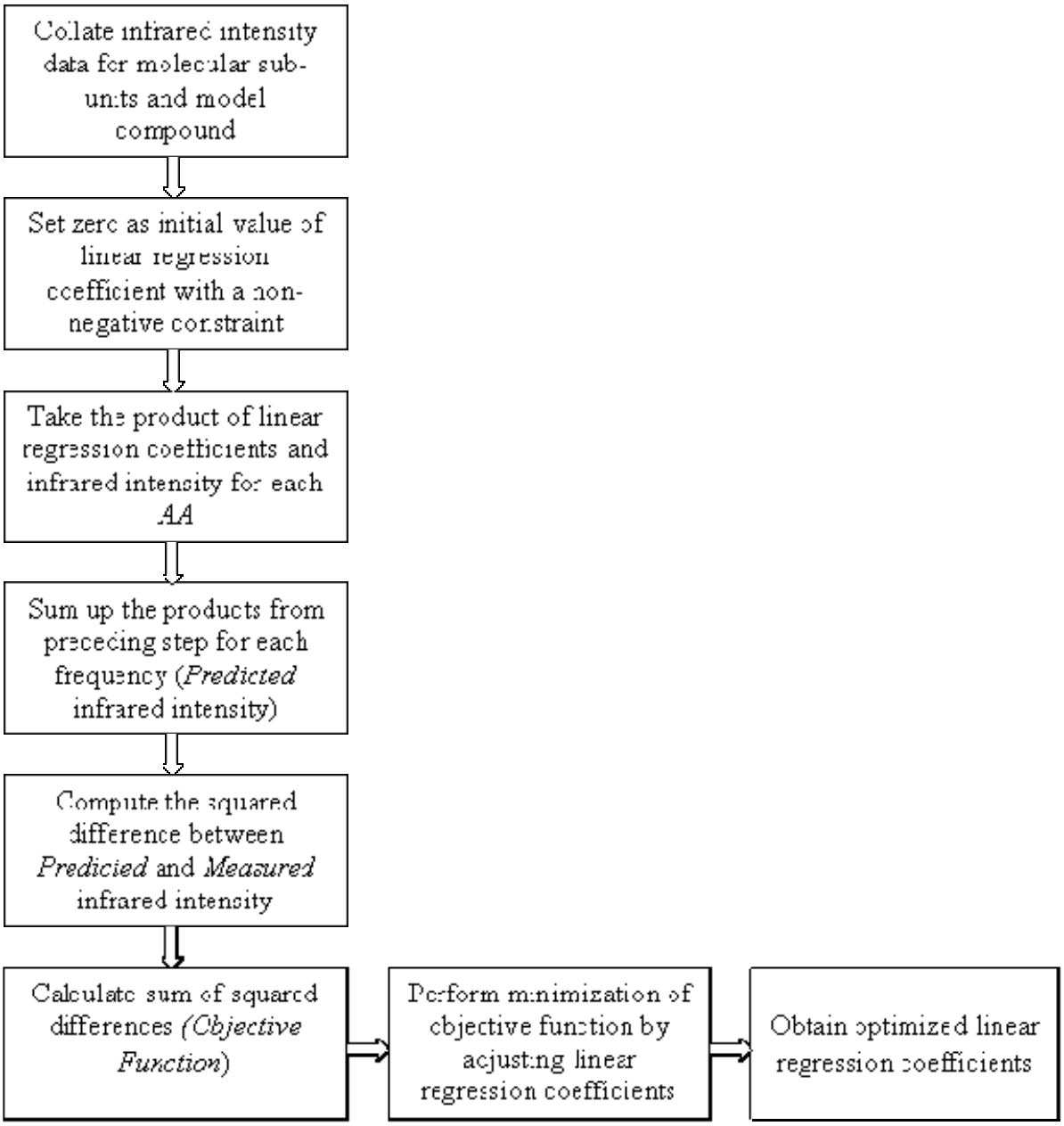


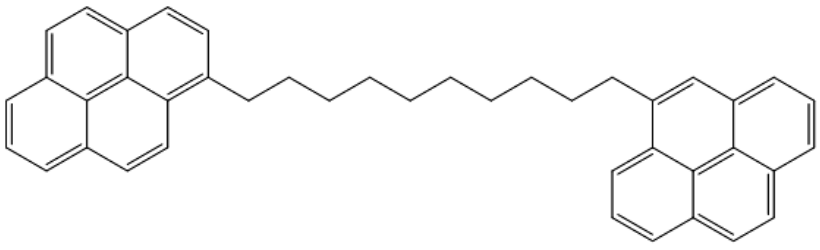
Figure 4-2: Schematic representation of optimization scheme employed in study (the process is repeated for Raman activity data)

As an illustrative example, *pyrene* +*decane*+ *pyrene* (Fig. 4-3) contains two pyrene rings connected by one alkane chain, thus the ratio of pyrene to decane is qualitatively 2:1. The infrared data of pyrene and decane (Figures 4-4(a) and 4-4(b) respectively) constitute the dependent variables, whose coefficients are to be determined, while *pyrene* + *decane* + *decane* represents the independent variable. The result of the regression analysis shows that the coefficient of pyrene is 1.56, while the coefficient of decane is 0.58. It is important to note that even though this ratio is greater than 2; it however shows that the coefficients provide a qualitative measure of the relative abundance of pyrene and decane in *pyrene* + *decane* + *pyrene.*

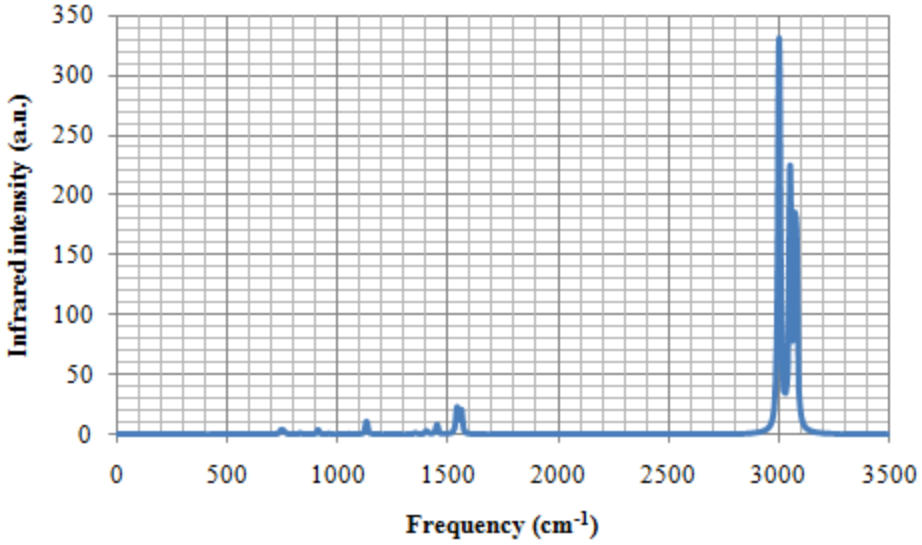
In Figure 4-4 (c), the difference between the computed (or calculated) spectrum of *pyrene* + *decane* +*pyrene* and the spectrum obtained after the spectral optimization i.e. the spectrum generated from the sum of 1.56 (pyrene) .and 0.58 (decane) is shown.

The prominent peaks in the residual spectrum can be attributed to the "extra" C-H vibrations. These "extra" vibrations can arise because of the four additional hydrogen atoms present in pyrene and decane, but absent in *pyrene* + *decane*+ *pyrene*. For the complete set of infrared and Raman results, see Appendix II.

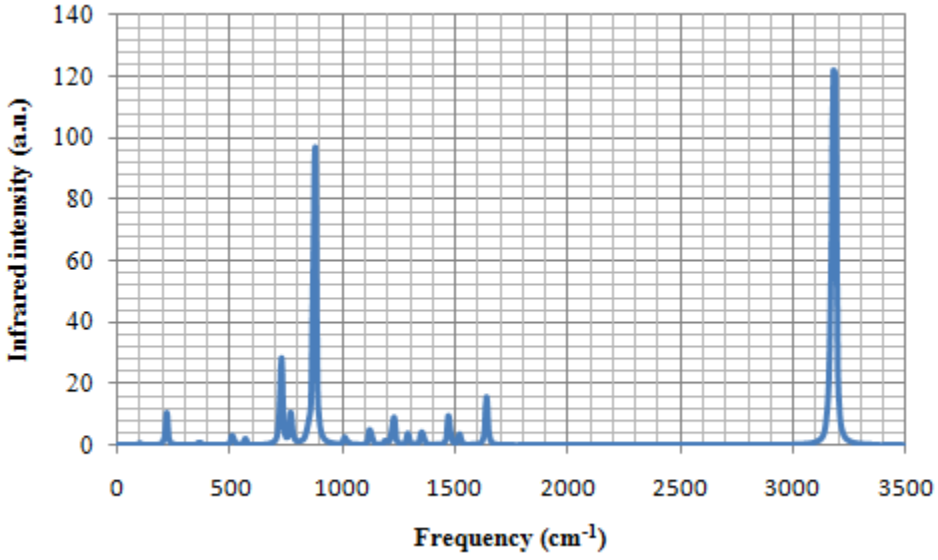
(a)



(b)


Infrared intensity (a.u.)
350
300
250
200
150
100
50
0
0
500
1000
1500
2000
2500
3000
3500
Frequency (cm-1)

(c)


Infrared intensity (a.u.)
140
120
100
80
60
40
20
0
0
500
1000
1500
2000
2500
3000
3500
Frequency (cm-1)

(d)

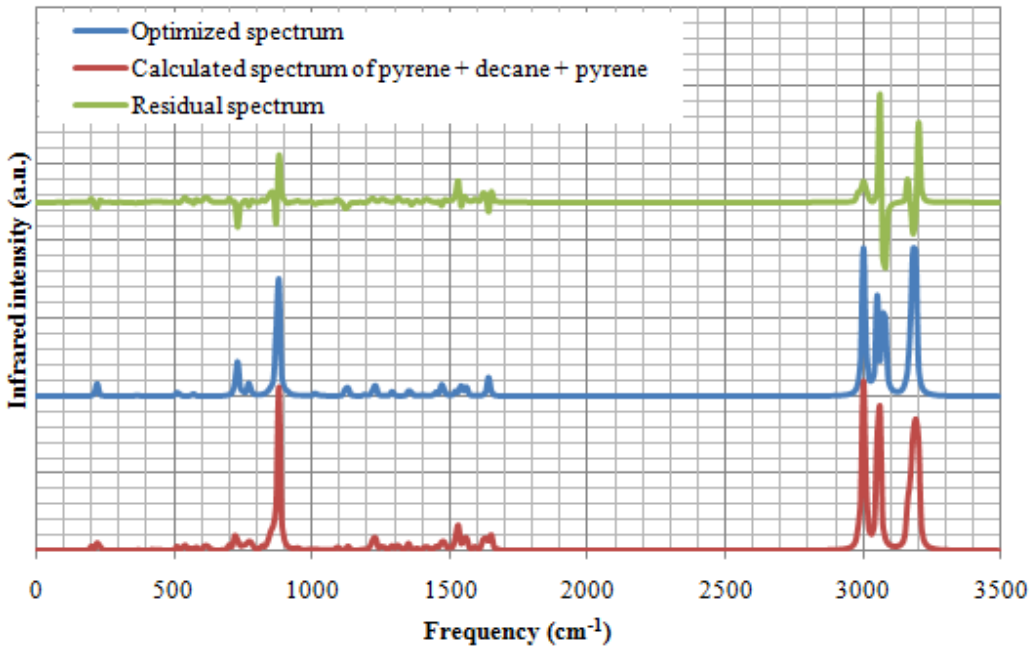


Figure 4-3: (a) *Pyrene* + *decane* + *pyrene*; the ratio of pyrene to decane is 2:1 (b)Infrared spectrum of decane (c) Infrared spectrum of pyrene (these two represent the dependent variables) (d) Infrared spectrum of *pyrene* + *decane* + *decane* (computed), spectrum obtained after the regression analysis (optimized), and the difference between the computed and optimized spectra (residual).

A more detailed description of the results obtained for all the model compounds studied using infrared and Raman spectra is provided in the following sections. The results obtained by analyzing the six model compounds consisting of identical aromatic rings connected by n-decane with a library of their constituent molecules are first presented; afterwards, the results from the entire database of model compounds analyzed with the complete library are discussed.

4.2 Infrared Spectra

Figures 4-4 and 4-5 show the results obtained for *pyrene* + *decane* + *pyrene* (*PDP*) and *anthracene* + *decane* + *anthracene* (*ADA*) using a library consisting of pyrene and decane, and anthracene and decane respectively. The coefficients

indicate that the ratio of pyrene to decane in *PDP* is approximately 2.70; while the ratio of anthracene to decane in *ADA* is approximately 1.94. Qualitatively, these results indicate that the contribution of pyrene and anthracene in *PDP* and *ADA* are at least twice the contribution of decane.

A summary of the coefficients for the six aromatic ring and n-decane model compound combinations is presented in Table 4-2. *Naphthalene* + *decane* + *naphthalene* has the lowest aromatic ring to n-alkane ratio of 1.02, while the ratio of 4.26 for *phenanthrene* + *decane* + *phenanthrene* is the highest. The mean ratio for all six cases is 2.39. This result is quite remarkable if the fact that there was no bias in the data acquisition or analysis. The results show that this technique can qualitatively predict the relative abundance of the molecular constituents of the hydrocarbon species under investigation.

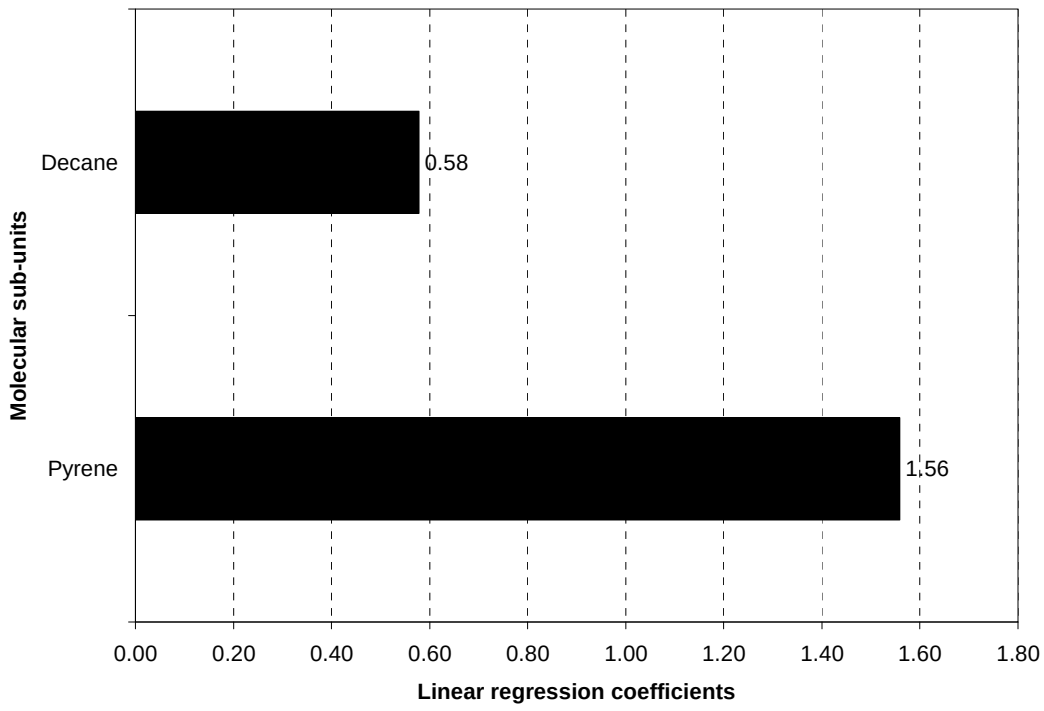


Figure 4-4: Linear regression coefficients for *pyrene* + *decane* + *pyrene* using a library containing pyrene and decane

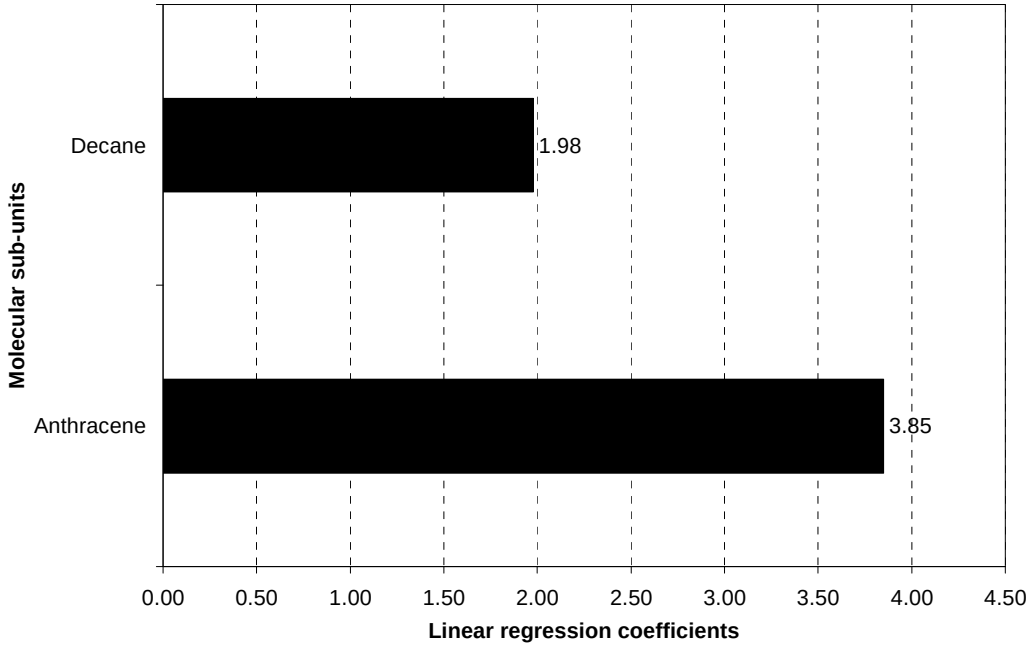


Figure 4-5: Linear regression coefficients for anthracene + decane + anthracene using a library consisting of anthracene and decane

Subsequently, the algorithm was applied to nine model compounds consisting of different aromatic rings connected by n-decane, using the complete library of six aromatic ring and five n-alkane chain compounds. The results of this exercise are reported in Table 4-4. The linear regression coefficient threshold of 0.3 was used to analyze the results. To illustrate the effect of using this approach, the result of *anthracene + decane + pyrene* is presented in Fig. 4-6. Even though anthracene, pyrene, and dodecane are correctly identified, the coefficient of 0.36 for phenanthrene indicates that the infrared result cannot be accepted as accurate.

Table 4-2: Infrared results for the six combinations of two identical aromatic rings joined by n-decane used to test the feasibility of using linear regression coefficient values in determining the relative abundance of molecular sub-units.

Model Compound	Molecular sub-unit (*Coefficient*)
anthracene + decane + anthracene	anthracene (*3.85*); decane (*1.98*)
chrysene + decane + chrysene	chrysene (*3.41*); decane (*1.55*)
fluoranthene + decane + phenanthrene	fluoranthene (*3.76*); decane (*1.72*)
naphthalene + decane + fluoranthene	naphthalene (*2.25*); decane (*2.21*)
phenanthrene + decane + pyrene	phenanthrene (*1.65*); decane (*0.39*)
pyrene + decane + pyrene	pyrene (*1.56*); decane (*0.58*)

The results showed a general trend: the identification of dodecane and/or undecane as the aliphatic chains present even though decane was the n-alkane chain used to construct the model compounds. To test the effect of alkane chain lengths an additional set of model compounds was developed that consisted of different combinations of pyrene and aromatic rings joined by varying even-numbered n-alkane chains lengths. The results of are presented in Table 4-5. The results show an approximately 40% success rate in identifying the molecular sub-units of the model compounds, but fails to distinguish between n-alkane chain lengths. Consequently, the identification of either undecane and/or dodecane has been chosen to represent the presence of an n-alkane chain.

Figure 4-7 is an illustration of the identification of dodecane instead of decane in *pyrene + decane + pyrene* using the complete library. The result indicates that pyrene is the most dominant structure in the model compound, followed by dodecane. This result reveals the strong potentials of using this technique to predict the molecular constituents of unknown hydrocarbon structures. *Pyrene + octane + pyrene* and *pyrene + dodecane + pyrene* both have coefficients of 0.31 for phenanthrene; however, because the corresponding values of pyrene are 1.28 and 1.41, the contribution of phenanthrene has been neglected.

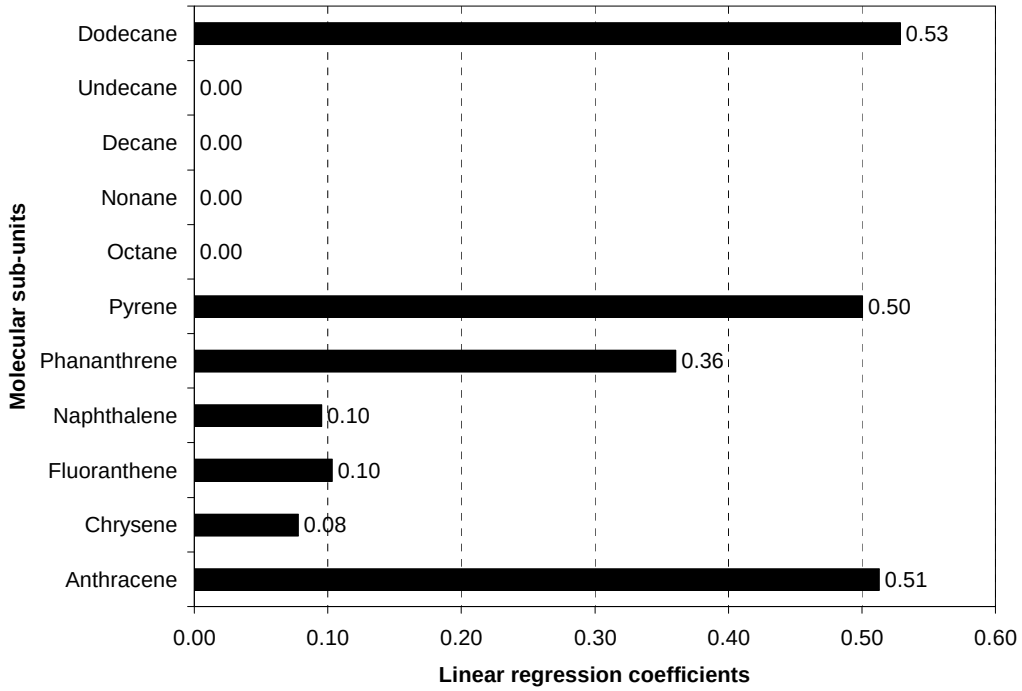


Figure 4-6: Infrared results for *anthracene* + *decane* + *pyrene* using a complete library of molecular sub-units.

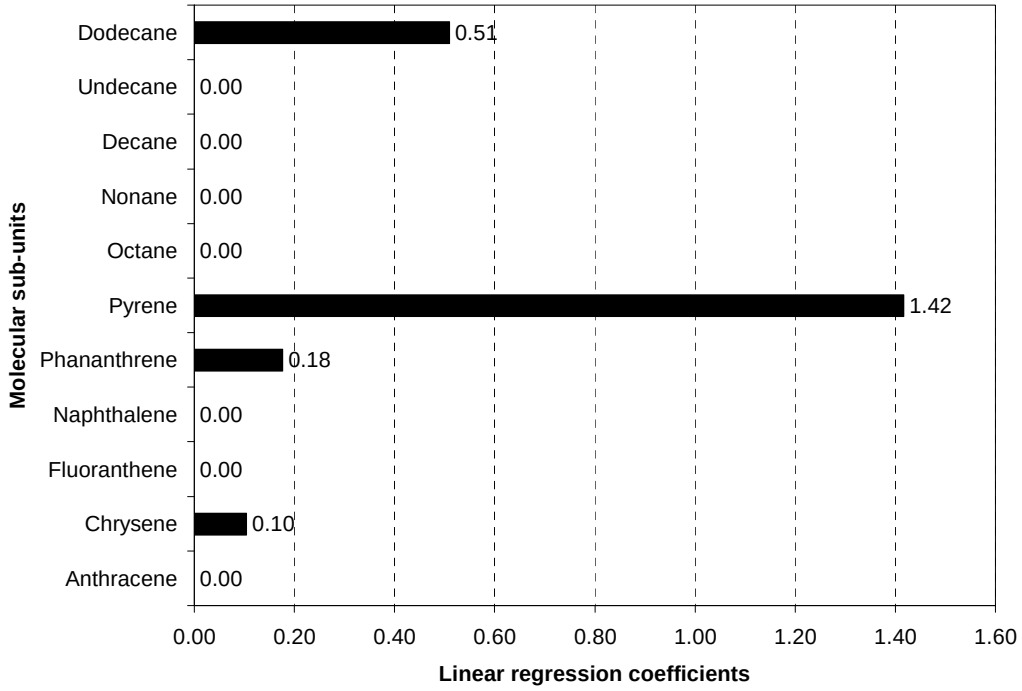


Figure 4-7: Infrared results for *pyrene* + *decane* + *pyrene* showing the coefficients of the complete library of molecular sub-units. Dodecane is identified instead on decane

The results for all three *anthracene*-based compounds show the correct identification of anthracene, pyrene and dodecane, but the misidentification of phenanthrene, which has an average coefficient of 0.37 for all cases (see Fig. 4-6); hence the technique fails to accurately identify this set of model compounds. However, it should be noted that the coefficients of anthracene, pyrene and dodecane are higher than the value for phenanthrene.

For the chrysene compound groups, the correct identification of the molecular sub-units is also achieved, with the misidentification of phenanthrene; however, unlike in the previous cases, the value of phenanthrene is higher than that of chrysene, but less than pyrene and dodecane for the two cases considered. Similarly, phenanthrene is misidentified in the naphthalene group of compounds, although it has a lower coefficient than naphthalene in all three cases.

In summary, the infrared results showed a 40% success rate in providing information regarding the aromatic ring and aliphatic chain compounds present in large molecules. While the technique fails to distinguish between aliphatic chains, it appears to misidentify phenanthrene sub units as anthracene, chrysene and naphthalene to varying degrees, Hence, the technique is unable to clearly identify the model compounds that they comprise.

4.3 Raman Spectra

Figure 4-8 shows the result obtained when the optimization algorithm is applied to *phenanthrene* + *decane* + *phenanthrene* using a library of phenanthrene and decane. The ratio of phenanthrene to decane is approximately 2.14; which evidently shows that the coefficients of these molecular sub-units provide a qualitative measure of their relative abundance in *phenanthrene* + *decane* + *phenanthrene*.

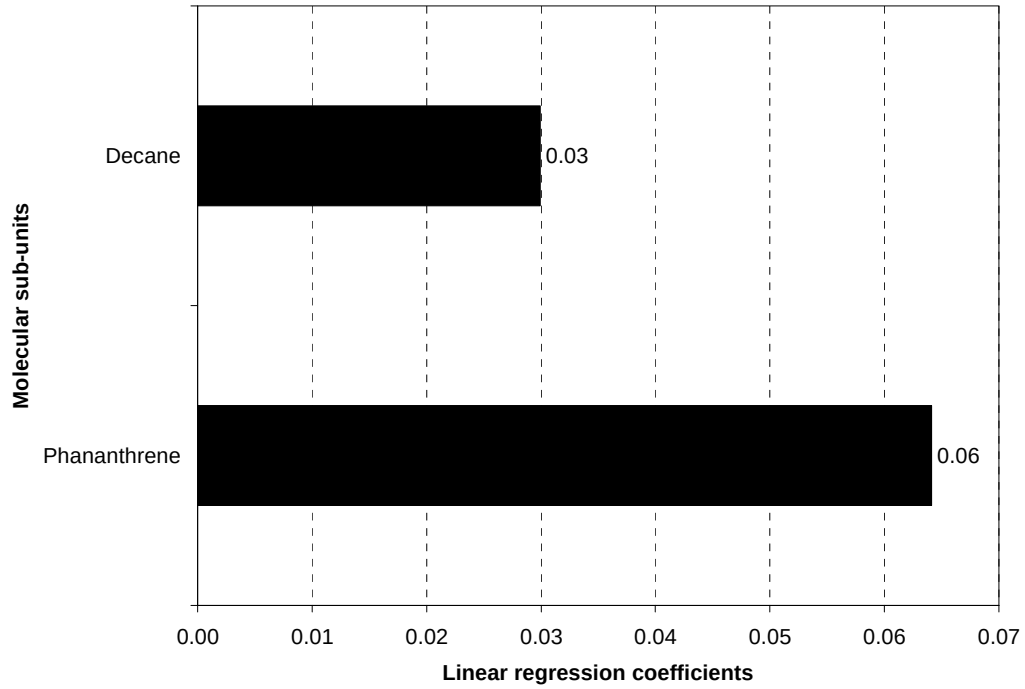


Figure 4-8: Raman results for *phenanthrene + decane + phenanthrene* using a library of its components

In Table 4-3, a summary of the results obtained for the six different combinations of two identical aromatic rings connected by n-decane model compounds is presented. The Raman results show a trend similar to the infrared case discussed in the preceding section: the aromatic rings have higher coefficients than n-decane. *Anthracene + decane + anthracene* has the lowest ratio of 1.37, while the ratio of 2.68 reported for *fluoranthene + decane + fluoranthene* is the highest. The average for all six compounds is approximately 1.79. Again, these values have been obtained without any bias in the data acquisition or analysis process.

Table 4-3: Raman results for the 6 combinations of two identical aromatic rings joined by n-decane used to test the feasibility of using linear regression coefficient values in determining the relative abundance of molecular sub-units

Model Compound	Molecular sub-unit (*Coefficient*)
anthracene + *decane* + *anthracene*	anthracene (*0.06*); decane (*0.04*)
chrysene + *decane* + *chrysene*	chrysene (*1.52*); decane (*1.17*)
fluoranthene + *decane* + *phenanthrene*	fluoranthene (*0.08*); decane (*0.03*)
naphthalene + *decane* + *fluoranthene*	naphthalene (*0.06*); decane (*0.04*)
phenanthrene + *decane* + *pyrene*	phenanthrene (*0.06*); decane (*0.03*)
pyrene + *decane* + *pyrene*	pyrene (*1.62*); decane (*1.05*)

As noted in the preceding sections, infrared and Raman data have been analyzed in exactly the same way. Hence, the next step in the analysis is the use of the complete library of eleven (six aromatic-ring and five n-alkane-chain) compounds. In Table 4-4, the Raman results for the nine model compounds comprising two different aromatic rings connected by n-decane are presented. Generally the results show a vast improvement over the infrared results. In the infrared case, only about 30% of the compounds were successfully identified, whereas the Raman results show almost 70 % success rate. Once again, the results reported are based on the compounds with regression coefficients above the threshold of 0.3

Table 4-5 shows the Raman results for the 14 compounds investigated to evaluate the ability of the technique to distinguish between different n-alkane chain-lengths. Unfortunately, the Raman results do not fare any better than the infrared ones, and n-alkane chain-lengths are still indistinguishable. Hence, the identification of any n-alkane chain has been chosen to represent the presence of a *bridge* linking the two aromatic rings

A close look at the results presented in Table 4-5 show that all of the *anthracene*-, *chrysene*-, and *phenanthrene*-based compounds are completely identified, while

none of the *naphthalene*-based ones are identified. The results for *naphthalene* + *octane* + *pyrene* and *pyrene* + *decane* + *pyrene* are reported in Figures 4-9 and 4-10 respectively. The results for all three *naphthalene*-based compounds show the misidentification of phenanthrene, while the unsuccessful *pyrene*-based compounds show a misidentification of naphthalene.

Comparing infrared and Raman results reveal their complementary nature. For instance, while the Raman results show a complete identification of all the *anthracene*-based model compounds, the infrared results fail to completely identify any of them. However, it has to be noted that both sets of results also show some similarity; they both fail to identify any of the *naphthalene*-based compounds, but completely identify the *phenanthrene*-based ones. Furthermore, the infrared results for the same set of compounds show a 40% success rate, and the Raman equivalent show a 60% success rate.

The complementary nature of both results, coupled with the fact that the infrared and Raman data are from the same compounds motivated the use of another level of analysis that combined both sets of results. The goal of this exercise was to seek an improvement in the level of accuracy obtained when both data sets were analyzed independently.

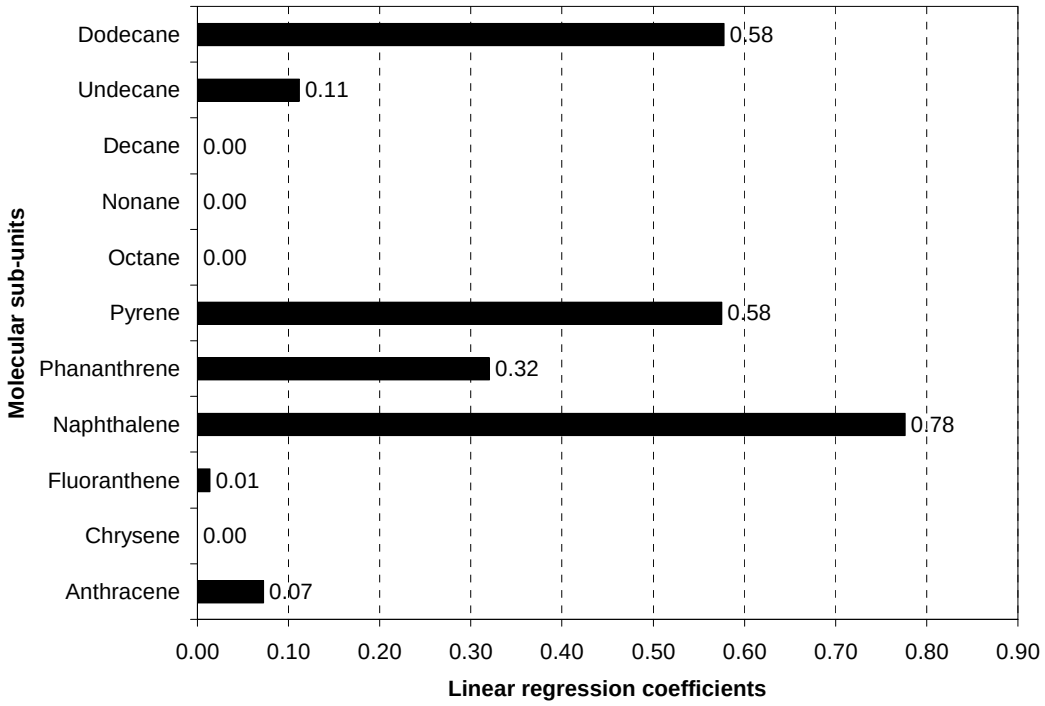


Figure 4-9: Raman result for *naphthalene + octane + pyrene* using the complete library of molecular sub-units

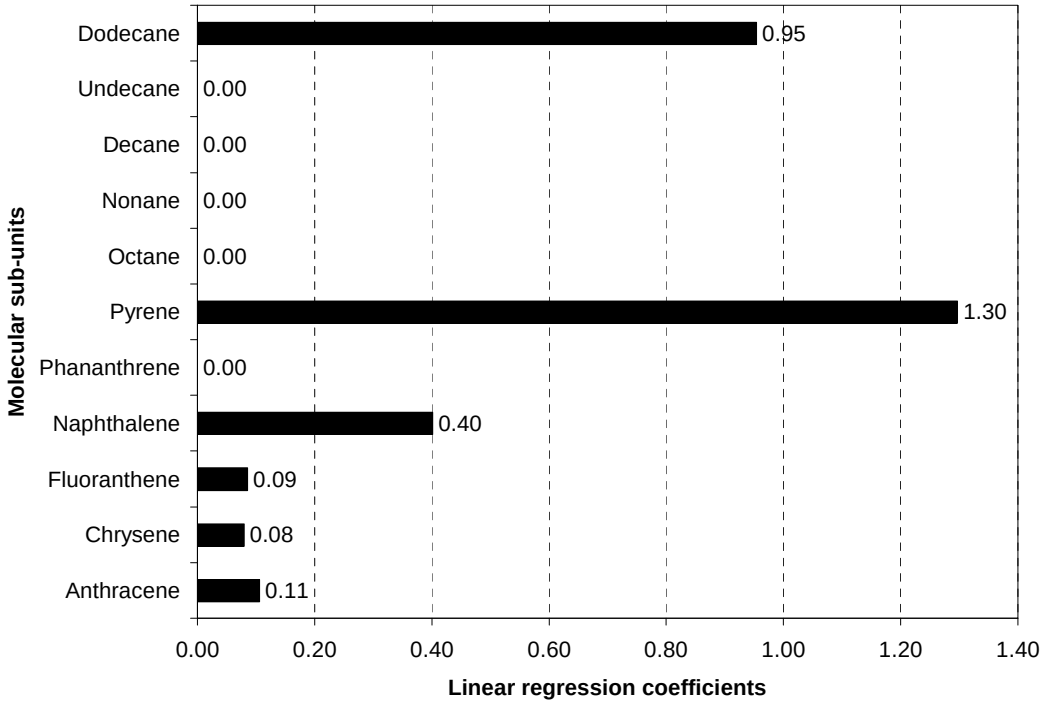


Figure 4-10: Raman results for *pyrene + decane + pyrene* using the complete library of molecular sub-units

4.4 Combined Infrared (IR) and Raman Spectroscopy

The combination of the infrared and Raman spectroscopy as a basis for molecule identification was motivated by the need to seek ways to improve on the accuracy of the optimization algorithm: 40 % using infrared spectroscopy and 65 % for Raman spectroscopy (based on the 20 model compounds investigated). An illustrative outcome is presented in Figure 4-11 where the infrared and Raman results for *anthracene* + *decane* + *pyrene* are plotted on the same chart. The jointly identified molecular sub-units of this compound (anthracene, dodecane and pyrene) are clear. This approach was applied to the entire database of model compounds, and the results are presented in Tables 4-4 and 4-5.

The combination of IR and Raman results produced an improvement in the prediction of the molecular sub-units for all the model compounds except for those containing *naphthalene* groups. As Figure 4-12 shows, the apparent presence of phenanthrene is once again responsible for this failure to identify compounds correctly.

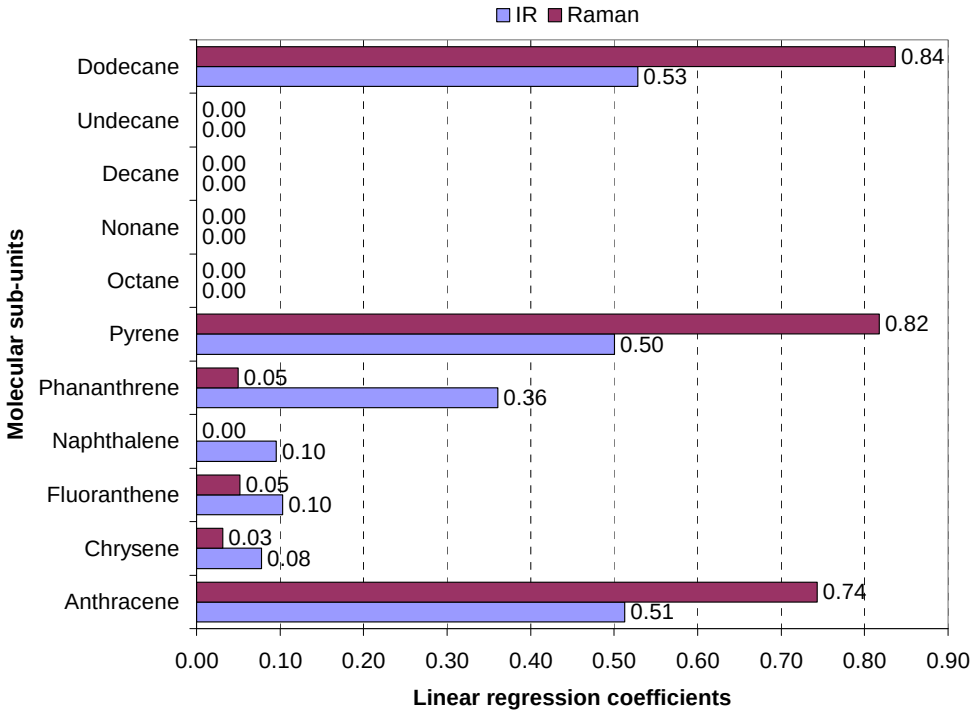


Figure 4-11: Combined IR and Raman result for *anthracene* + *decane* + *pyrene* using the complete library of molecular sub-units

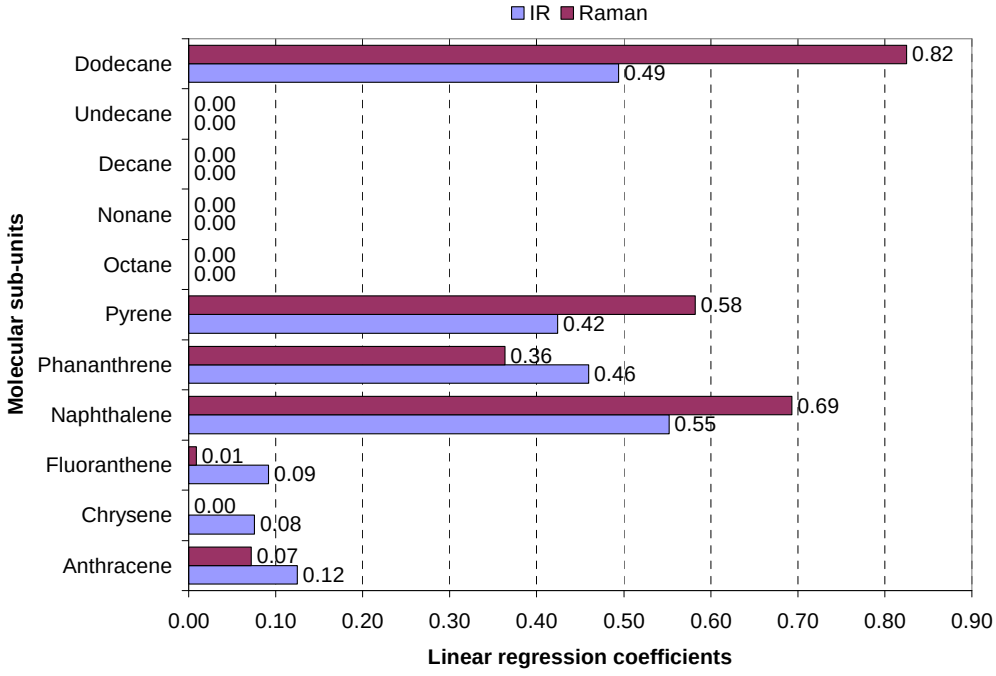


Figure 4-12: Combined infrared and Raman result for *naphthalene* + *decane* + *pyrene* using the complete library of molecular sub-units

Table 4-4: IR, Raman, and combined IR and Raman results for the 9 model compounds consisting of different combinations of two aromatic rings connected by an n-decane chain (✓ = Pass; ✖ = Fail)

Model Compound	IR	Raman	IR + Raman
anthracene + decane + chrysene	✖	✓	✓
anthracene + decane + pyrene	✖	✓	✓
chrysene + decane + fluoranthene	✓	✓	✓
chrysene + decane + phenanthrene	✓	✓	✓
fluoranthene + decane + phenanthrene	✖	✓	✓
naphthalene + decane + fluoranthene	✖	✖	✖
naphthalene + decane + pyrene	✖	✖	✖
naphthalene + decane + phenanthrene	✖	✖	✖
phenanthrene + decane + pyrene	✓	✓	✓

Table 4-5: Summary of the IR, Raman, and combined IR and Raman results for the 14 model compounds used to evaluate the influence of n-alkane chain lengths on the optimization technique. (✓ = Pass; ✖ = Fail)

Model Compound	IR	Raman	IR + Raman
anthracene + octane + pyrene	✖	✓	✓
anthracene + decane + pyrene	✖	✓	✓
anthracene + dodecane + pyrene	✖	✓	✓
pyrene + octane +pyrene	✓	✖	✓
pyrene + decane +pyrene	✓	✖	✓
pyrene + dodecane +pyrene	✓	✓	✓
chrysene + octane + pyrene	✖	✓	✓
chrysene + dodecane + pyrene	✖	✓	✓
naphthalene + octane + pyrene	✖	✖	✖
naphthalene + decane + pyrene	✖	✖	✖
naphthalene + dodecane + pyrene	✖	✖	✖
phenanthrene + octane + pyrene	✓	✓	✓
phenanthrene + decane + pyrene	✓	✓	✓
phenanthrene + dodecane + pyrene	✓	✓	✓

4.5 Compounds with no Library Constituents

The results reported in the preceding sections are based on model compounds constructed from the molecular sub-units that make-up the library. Although the results are very promising, it is still pertinent to consider a situation where the model compound is not composed of any of the molecular sub-units in the library. This test is important because a positive identification of the model compound would indicate that the spectral optimization process that forms the basis for this thesis is unreliable; whereas a negative identification would further strengthen the case for molecular structure prediction using this technique.

Figures 4-13 and 4-14 are the schematic representation of the molecular structure of *azuelene* + *decane* + *fluorene*, and the combined IR and Raman result respectively. Using 0.3 as the threshold of significance for the coefficients, the IR result indicates that pyrene and dodecane are identified, while fluoranthene and decane are identified in the Raman. However, combining both results reveals that none of the molecular sub-units are jointly identified.

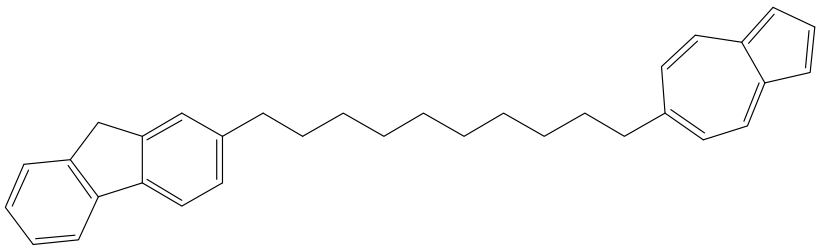


Figure 4-13: Schematic representation of *azulene + decane + fluorene*

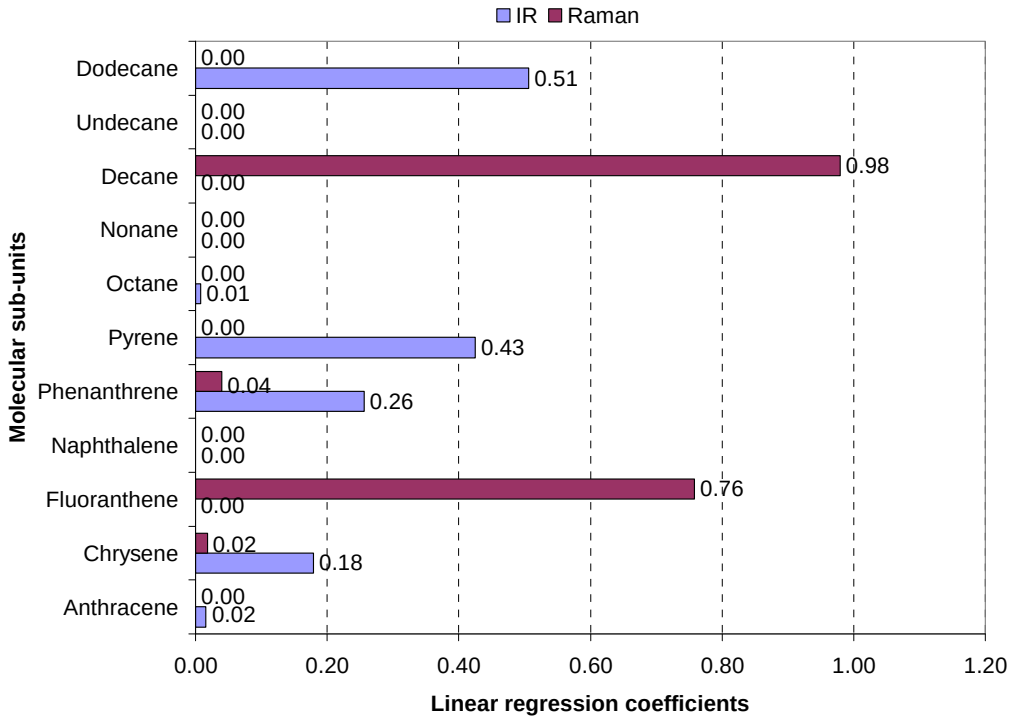


Figure 4-14: Combined IR and Raman result for *azulene* + *decane* + *fluorene* using the complete library of molecular sub-units

4.6 Analysis of Residuals

The spectrum that results from the difference between the computed (or measured) spectrum and optimized (or predicted) spectrum is the residual spectrum. An analysis of the residuals shows that on the average they represent 25% and 17% of the computed spectra for IR and Raman respectively. Further, it was observed that the size of the residuals was independent of the successful identification of the molecular sub-units.

For instance, for the three pyrene-based groups successfully identified in the IR analysis, the residuals represent 17%, 19%, and 19% for *pyrene* + *octane* + *pyrene, pyrene* + *decane* + *pyrene, and pyrene* + *dodecane* + *pyrene* respectively; while for the three naphthalene-based compounds which were unsuccessful namely *naphthalene* + *octane* + *pyrene, naphthalene* + *decane* +

pyrene, and naphthalene + dodecane + pyrene, the residuals represent 16%, 18%, and 18% respectively. A complete set of the results is provided in Appendix I.

4.7 Summary

The linear regression of the infrared and Raman data of 20 model compounds using a library of molecular sub-units was investigated. The linear regression coefficients obtained from the minimization process were used to provide an indication of the relative abundance of molecular sub-units in the model compounds. For compounds correctly identified by way of constituents, average compositions are semi quantitative. However the variation among specific cases is significant. 40% of the compounds were successfully identified, on the basis of IR spectra, while 65 % were correctly identified on the basis of Raman spectra. By combining these two sets of independently obtained results more than 80% of compounds were identified correctly, and the variation of relative abundance of sub units was reduced. Alkane chain length is poorly defined.

5 Nuclear Magnetic Resonance (NMR)

5.1 Introduction

As noted in Chapter 2, advanced stochastic algorithms such as Monte Carlo simulation have been used to generate molecular representations of asphaltenes based on their experimental NMR (^{1}H and ^{13}C) spectroscopy, elemental analysis and molecular weight data. However, based on the results reported by Sheremata *et al.* [43] and Boek *at al* [44], the molecular structure of heavy petroleum fractions such as asphaltenes remains ambiguous. While Sheremata *et al.* concluded that asphaltenes had an archipelago representation; Boek *et al.* were able to show that the polycondensed representation was more accurate. It should be noted that Boek *et al.* used 30 aromatic building blocks, which is 10 more than Sheremata *et al.'s*; and the authors also claim that unlike Sheremata *et al.*, they did not constrain their results to the archipelago structure.

Following the successful application of the spectral optimization algorithm to the vibrational spectra of the model compounds investigated; the ambiguity generated in the aforementioned publications motivated the extension of this algorithm to NMR spectra. For consistency with infrared and Raman results, the NMR data used for this analysis was obtained using exactly the same DFT level of theory (B3LYP/6-311G).

5.2 NMR Results

Unlike the vibrational frequency data, the NMR data is characterized by Chemical Shift or Shielding Tensor measured in parts per million (ppm) and Degeneracy (which represents the number of atoms (Carbon for ^{13}C-NMR, or Hydrogen for ^{1}H-NMR) with the same Chemical Shift or Shielding Tensor. For ease of analysis, the Shielding Tensors have been used instead of the Chemical Shift values.

Tables 5-1 and 5-2 show the results from the analysis conducted using the technique outlined in Chapter 4. None of the model compounds were correctly identified. Figures 5-1 and 5-2 depict the ^{13}C-NMR result for *pyrene* + *octane* + *pyrene* and the ^{1}H-NMR result for *anthracene* + *octane* + *pyrene* respectively. These results are typical for all the model compounds studied. A complete set of NMR results comprise Appendix III.

Table 5-1: ^{1}H and ^{13}C-NMR results for the 9 model compounds consisting of different combinations of two aromatic rings connected by an n-decane chain (✓ = Pass; ✖ = Fail)

Model Compound	^{1}H-NMR	^{13}C-NMR
anthracene + *decane* + *chrysene*	✖	✖
anthracene + *decane* + *pyrene*	✖	✖
chrysene + *decane* + *fluoranthene*	✖	✖
chrysene + *decane* + *phenanthrene*	✖	✖
fluoranthene + *decane* + *phenanthrene*	✖	✖
naphthalene + *decane* + *fluoranthene*	✖	✖
naphthalene + *decane* + *pyrene*	✖	✖
naphthalene + *decane* + *phenanthrene*	✖	✖
phenanthrene + *decane* + *pyrene*	✖	✖

Table 5-2: Summary of the ^{1}H and ^{13}C-NMR results for the 14 model compounds used to evaluate the influence of n-alkane chain lengths on the optimization technique. (✓ = Pass; ✖ = Fail)

Model Compound	^{1}H –NMR	^{13}C-NMR
anthracene + octane + pyrene	✖	✖
anthracene + decane + pyrene	✖	✖
anthracene + dodecane + pyrene	✖	✖
pyrene + octane +pyrene	✖	✖
pyrene + decane +pyrene	✖	✖
pyrene + dodecane +pyrene	✖	✖
chrysene + octane + pyrene	✖	✖
chrysene + dodecane + pyrene	✖	✖
naphthalene + octane + pyrene	✖	✖
naphthalene + decane + pyrene	✖	✖
naphthalene + dodecane + pyrene	✖	✖
phenanthrene + octane + pyrene	✖	✖
phenanthrene + decane + pyrene	✖	✖
phenanthrene + dodecane + pyrene	✖	✖

These results show that irrespective of the model compound used, any combination of the aromatic rings can be predicted. Although these results suggest that the optimization algorithm that was successful with vibrational spectra has completely failed with NMR spectra; it however corroborates the findings of Boek *et al.* and Sheremata *et al.*: the same NMR data can be used to generate a plethora of molecular representations. Further, since the electron densities of model compounds studied are very similar, their NMR signatures are very nearly the same; this could also account for the lack of resolution observed using NMR.

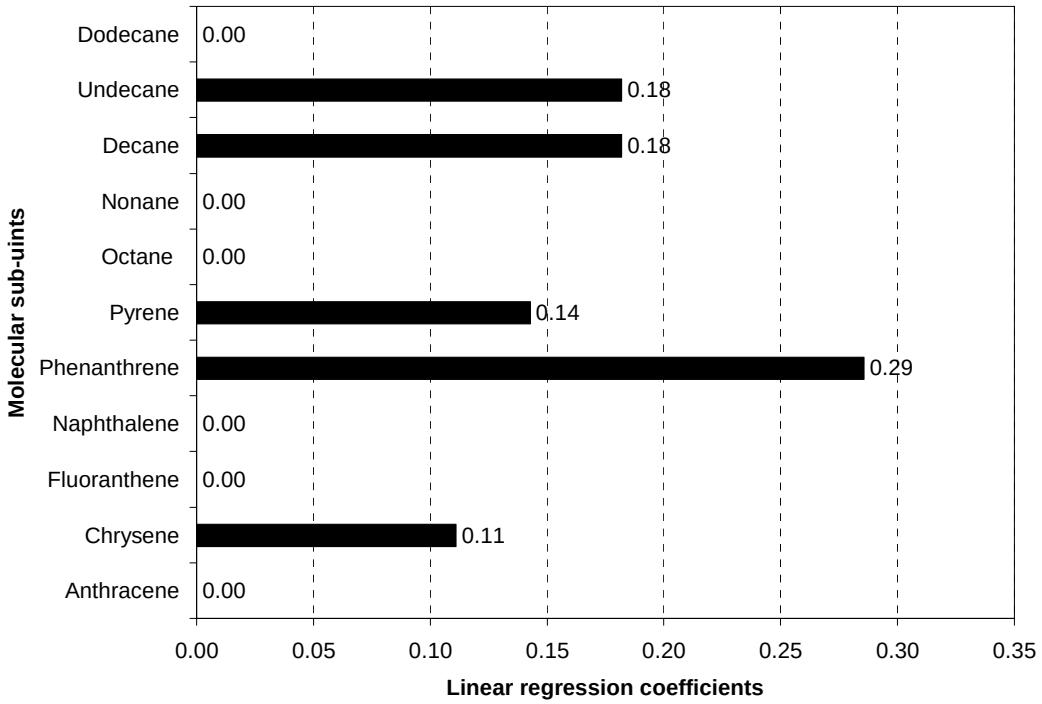


Figure 5-1: ^{13}C NMR Result for *pyrene* + *octane* + *pyrene* using the complete library of molecular sub-units

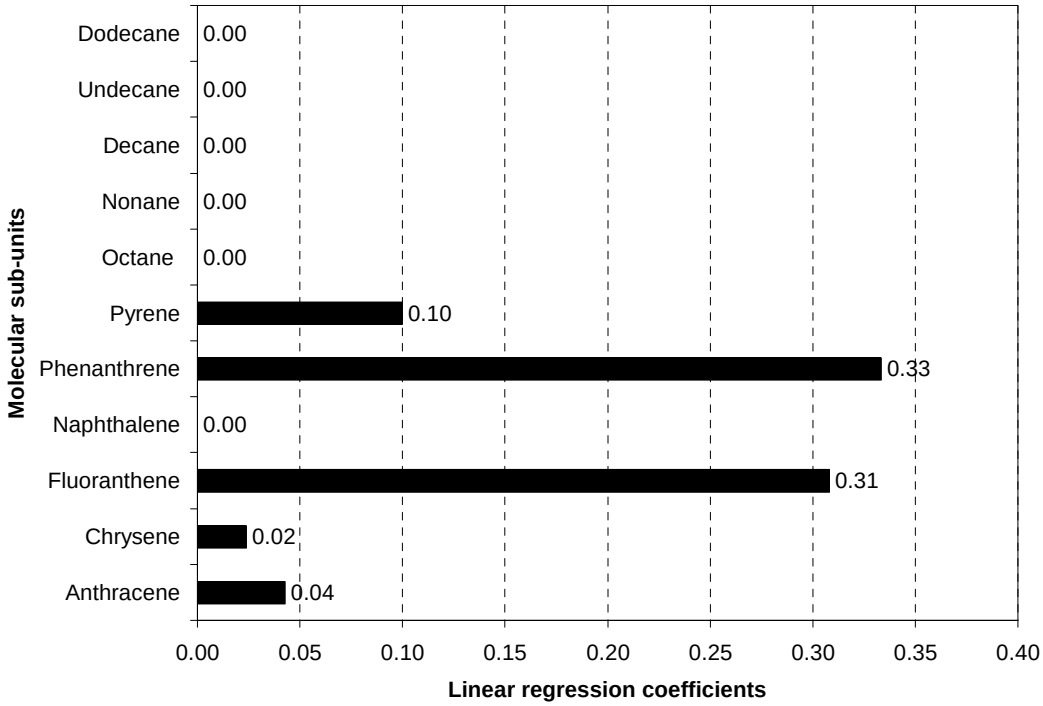


Figure 5-2: ^{1}H-NMR result for *anthracene* + *octane* + *pyrene* using the complete library of molecular sub-units

6 Conclusion

6.1 Introduction

This thesis was motivated by the need to reduce the ambiguity of molecular representations proposed for ill-defined petroleum fractions including asphaltenes, vacuum residues, heavy oils and bitumen. A review of the literature showed that the use of vibrational spectroscopy has played a minor role in this debate, whereas techniques that rely on Nuclear Magnetic Resonance (NMR) have been heavily relied upon.

An optimization technique was developed to explore the feasibility of using vibrational frequency data to predict the molecular structure of this class of petroleum fractions, and the results are promising. Since this study is exploratory in nature, computed spectra were employed.

6.2 Main findings

The results presented in the preceding chapters support the use of infrared and Raman spectroscopy in predicting the molecular structure of hydrocarbon structures; and this can be extended to complex, ill-defined hydrocarbons including asphaltenes, vacuum residue, and bitumen.

Using infrared data alone, approximately 40% of the model compounds studied was completely identified; likewise, the use of Raman data alone resulted in 65% identification. The combination of these two methods further boosted the accuracy of the technique, leading to more than 80% of the model compounds being completely identified.

By contrast, a library of NMR spectra (^{1}H and ^{13}C) for small molecules failed to identify large model compounds correctly; the results indicate that one set of NMR data can be used to construct diverse structures for larger hydrocarbon molecules.

The NMR spectrum of a hydrocarbon compound only provides information about the individual carbon and hydrogen atoms, and their nearest neighbours; whereas the vibrational spectroscopy data of the same compound shows the functional groups present. These functional groups represent a collection of carbon and hydrogen atoms, rather than individual atoms; therefore the functional groups provide a clearer picture of the molecular sub-units present.

Spectral optimization has been used to successfully predict the molecular constituents of large model hydrocarbon molecules based on infrared and Raman spectroscopy data. Infrared and Raman spectroscopy can therefore play an important role in resolving the ambiguity surrounding the molecular structure of ill-defined hydrocarbons.

6.3 Recommendations

The findings presented in this work are based on molecular simulations of compounds in their ground states and vapour phase. A study based on experimental spectroscopic data of the same set of model compounds followed by one including bitumen, asphaltenes, and heavy oils is warranted. Experimental work in this area presents numerous challenges. For example, to obtain adequate resolution of IR spectra, particularly at low wave numbers, synchrotron based experiments will be required and specialized experimental techniques may well be required (e.g.: photoacoustic IR). Raman measurements may be performed using typical laboratory equipment but samples may well need to be cooled, and care must be taken with the selection of laser wavelength.

An additional computational study warranted on the basis of the results reported here concerns whether IR and RAMAN spectra can discriminate naphthenic or partially naphthenic subunits from corresponding polynuclear aromatic ones. This remains a challenging experimental area for other forms of spectroscopy.

References

1. Cosio, M. G.;Saetta, M., and Agusti, A., *Immunologic Aspects of Chronic Obstructive Pulmonary Disease.* N Engl J Med, 2009. **360**(23): p. 2445-2454.

2. Wiehe, I. A. and Liang, K. S., *Asphaltenes, Resins, and Other Petroleum Macromolecules.* Fluid Phase Equilibria, 1996. **117**(1-2): p. 201-210.

3. Speight, J. G., *The Chemistry and Technology of Petroleum*. 4th ed. ed. 2007, Florida: CRC Press.

4. Marshall, A. G. and Rodgers, R. P.,*Petroleomics: The Next Grand Challenge for Chemical Analysis*. 2004. p. 53-59.

5. Head, I. M.;Jones, D. M., and Larter, S. R., *Biological Activity in the Deep Subsurface and the Origin of Heavy Oil.* Nature, 2003. **426**(6964): p. 344-352.

6. R. F. Meyer, E. D. A., (2003) *Heavy Oil and Natural Bitumen-Strategic Petroleum Resources: U.S. Geological Survey Fact Sheet 70-03, Available Online at Http://Pubs.Usgs.Gov/Fs/Fs070-03/Fs070-03.Html. Accessed May 13, 2009*.

7. Leon, V. and Kumar, M., *Biological Upgrading of Heavy Crude Oil.* Biotechnology and Bioprocess Engineering, 2005. **10**(6): p. 471-481.

8. Nji, G. N.;Svrcek, W. Y.;Yarranton, H. W., and Satyro, M. A.,*Characterization of Heavy Oils and Bitumens. 1. Vapor Pressure and Critical Constant Prediction Method for Heavy Hydrocarbons*. 2008. p. 455-462.

9. Coniglio, L.;Trassy, L., and Rauzy, E.,*Estimation of Thermophysical Properties of Heavy Hydrocarbons through a Group Contribution Based Equation of State*. 2000. p. 5037-5048.

10. Calles, A. V. W.,University of Alberta, *Estimation of Thermophysical Properties of Athabasaca Vacuum Residue Using a Group Contribution Based Equation of State*, in *Chemical Engineering*. 2006: Edmonton.

11. Saber, N. and Shaw, J. M., *Toward Multiphase Equilibrium Prediction for Ill-Defined Asymmetric Hydrocarbon Mixtures.* Fluid Phase Equilibria, 2009. **285**(1-2): p. 73-82.

12. Marrero, J. and Gani, R., *Group-Contribution Based Estimation of Pure Component Properties.* Fluid Phase Equilibria, 2001. **183-184**: p. 183-208.

13. Sallamie, N. and Shaw, J. M., *Heat Capacity Prediction for Polynuclear Aromatic Solids Using Vibration Spectra.* Fluid Phase Equilibria, 2005. **237**(1-2): p. 100-110.

14. Lastovka, V.;Sallamie, N., and Shaw, J. M., *A Similarity Variable for Estimating the Heat Capacity of Solid Organic Compounds: Part I. Fundamentals.* Fluid Phase Equilibria, 2008. **268**(1-2): p. 51-60.

15. Lastovka, V.;Fulem, M.;Becerra, M., and Shaw, J. M., *A Similarity Variable for Estimating the Heat Capacity of Solid Organic Compounds: Part Ii. Application: Heat Capacity Calculation for Ill-Defined Organic Solids.* Fluid Phase Equilibria, 2008. **268**(1-2): p. 134-141.

16. Lastovka, V. and Shaw, J. M.,*Predictive Correlation for Cp of Organic Solids Based on Elemental Composition*. 2007. p. 1160-1164.

17. Lastovka, V. and Shaw, J. M., *Overview of Heat Capacity Models for Hydrocarbons and a Novel Ideal Gas Model, Petroleum Thermodynamics Research Chair Advisory Committee Meeting November 13th, 2008*

18. Speight, J. G.,*Petroleum Asphaltenes. Part 1 : Asphaltenes, Resins and the Structure of Petroleum*. 2004. p. 467-477.

19. Liu, Y. C.;Sheu, E. Y.;Chen, S. H., and Storm, D. A., *Fractal Structure of Asphaltenes in Toluene.* Fuel, 1995. **74**(9): p. 1352-1356.

20. Maham, Y.;Chodakowski, M. G.;Zhang, X., and Shaw, J. M., *Asphaltene Phase Behavior: Prediction at a Crossroads.* Fluid Phase Equilibria, 2005. **228-229**: p. 21-26.

21. Behar, F.;Pelet, R., and Roucache, J., *Geochemistry of Asphaltenes.* Organic Geochemistry, 1984. **6**: p. 587-595.

22. Marshall, A. G. and Rodgers, R. P., *Petroleomics: The Next Grand Challenge for Chemical Analysis.* Accounts of Chemical Research, 2004. **37**(1): p. 53-59.

23. Strausz, O. P.;Lown, E. M., and Mojelsky, T. W. *Chemical Characterization of Alberta Oil Sand Bitumen*. in *5th UNITAR International Conference of Heavy Crude and Tar Sands*. 1991.

24. Jaffe, S. B.;Freund, H., and Olmstead, W. N.,*Extension of Structure-Oriented Lumping to Vacuum Residua*. 2005. p. 9840-9852.

25. Walters, C., *The Origin of Petroleum*, in *Practical Advances in Petroleum Processing*. 2006. p. 79-101.

26. Robinson, R., *The Origins of Petroleum.* Nature, 1966. **212**(5068): p. 1291-1295.

27. Durand, B.,Oil & Gas Science and Technology - Rev. IFP, *A History of Organic Geochemistry* 2003. p. 203-231.

28. Oakwood, T. S.;Shriver, D. S.;Fall, H. H.;McAleer, W. J., and Wunz, P. R.,*Optical Activity of Petroleum*. 1952. p. 2568-2570.

29. Hunt, J. M., *Petroleum Geochemistry and Geology*. 2nd ed. ed. 1996, New York: W. H. Freeman.

30. Chosson, P.;Lanau, C.;Connan, J., and Dessort, D., *Biodegradation of Refractory Hydrocarbon Biomarkers from Petroleum under Laboratory Conditions.* Nature, 1991. **351**(6328): p. 640-642.

31. Meyer, R. F.;Attanasi, E. D., and Freeman, P. A., (2007) *Heavy Oil and Natural Bitumen Resources in Geological Basins of the World: U.S. Geological Survey Open File Report 2007-1084, Available Online at Http://Pubs.Usgs.Gov/of/2007/1084/; Accessed May 13, 2009*.

32. Miller, J. T.;Fisher, R. B.;Thiyagarajan, P.;Winans, R. E., and Hunt, J. E., *Subfractionation and Characterization of Mayan Asphaltene.* Energy & Fuels, 1998. **12**(6): p. 1290-1298.

33. Tissot, B. P. and Welte, D. H., *Petroleum Formation and Occurence*. 2nd ed. ed. 1984.

34. Altget, K. H. and Boduszynski, M. M., *Composition and Analysis of Heavy Petroleum Fractions*. 1994.

35. Hannisdal, A.;Hemmingsen, P. V., and Sjoblom, J.,*Group-Type Analysis of Heavy Crude Oils Using Vibrational Spectroscopy in Combination with Multivariate Analysis*. 2005. p. 1349-1357.

36. Pelet, R.;Behar, F., and Monin, J. C., *Resins and Asphaltenes in the Generation and Migration of Petroleum.* Organic Geochemistry, 1986. **10**(1-3): p. 481-498.

37. Strausz, O. P.;Mojelsky, T. W.;Faraji, F.;Lown, E. M., and Peng, P. a.,*Additional Structural Details on Athabasca Asphaltene and Their Ramifications*. 1999. p. 207-227.

38. Sheu, E. Y.,*Petroleum Asphaltene Properties, Characterization, and Issues*. 2002. p. 74-82.

39. Strausz, O. P.;Mojelsky, T. W.;Lown, E. M.;Kowalewski, I., and Behar, F.,*Structural Features of Boscan and Duri Asphaltenes*. 1999. p. 228-247.

40. Payzant, J. D.;Lown, E. M., and Strausz, O. P., *Structural Units of Athabasca Asphaltene: The Aromatics with a Linear Carbon Framework.* Energy & Fuels, 1991. **5**(3): p. 445-453.

41. Yen, T. F.,Taylor & Francis, *Structure of Petroleum Asphaltene and Its Significance*. 1974. p. 447 - 463.

42. Zhao, S.;Kotlyar, L. S.;Woods, J. R.;Sparks, B. D.;Hardacre, K., and Chung, K. H., *Molecular Transformation of Athabasca Bitumen End-Cuts During Coking and Hydrocracking.* Fuel, 2001. **80**(8): p. 1155-1163.

43. Sheremata, J. M.;Gray, M. R.;Dettman, H. D., and McCaffrey, W. C., *Quantitative Molecular Representation and Sequential Optimization of Athabasca Asphaltenes.* Energy and Fuels, 2004. **18**(5): p. 1377-1384.

44. Boek, E. S.;Yakovlev, D. S., and Headen, T. F., *Quantitative Molecular Representation of Asphaltenes and Molecular Dynamics Simulation of Their Aggregationâ€* Energy & Fuels, 2009. **23**(3): p. 1209-1219.

45. Balci, M., *Basic 1h and 13c Spectroscopy*. 2005: Elsevier.

46. Sheremata, J. M.,University of Alberta, *Residue Molecules: Molecular Representations and Thermal Reactivity Phd Thesis*, in *Chemical and Materials Engineering*. 2008: Edmonton.

47. Yen, T. F.;Wu, W. H., and Chilingar, G. V., *A Study of the Structure of Petroleum Asphaltenes and Related Substances by Proton Nuclear Magnetic Resonance.* Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 1984. **7**(3): p. 275 — 304.

48. Yen, T. F.;Erdman, J. G., and Pollack, S. S.,*Investigation of the Structure of Petroleum Asphaltenes by X-Ray Diffraction*. 1961. p. 1587-1594.

49. Dickie, J. P. and Yen, T. F.,*Macrostructures of the Asphaltic Fractions by Various Instrumental Methods*. 1967. p. 1847-1852.

50. Shirokoff, J. W.;Siddiqui, M. N., and Ali, M. F.,*Characterization of the Structure of Saudi Crude Asphaltenes by X-Ray Diffraction*. 1997. p. 561-565.

51. Andersen, S. I.;Jensen, J. O., and Speight, J. G.,*X-Ray Diffraction of Subfractions of Petroleum Asphaltenes*. 2005. p. 2371-2377.

52. Groenzin, H. and Mullins, O. C., *Asphaltene Molecular Size and Structure.* The Journal of Physical Chemistry A, 1999. **103**(50): p. 11237-11245.

53. Andrews, A. B.;Guerra, R. E.;Mullins, O. C., and Sen, P. N., *Diffusivity of Asphaltene Molecules by Fluorescence Correlation Spectroscopy.* The Journal of Physical Chemistry A, 2006. **110**(26): p. 8093-8097.

54. Groenzin, H. and Mullins, O. C., *Molecular Size and Structure of Asphaltenes from Various Sources.* Energy and Fuels, 2000. **14**(3): p. 677-684.

55. Strausz, O. P.;Safarik, I.;Lown, E. M., and Morales-Izquierdo, A., *A Critique of Asphaltene Fluorescence Decay and Depolarization-Based Claims About Molecular Weight and Molecular Architecture.* Energy & Fuels, 2008. **22**(2): p. 1156-1166.

56. Mullins, O. C.,*Rebuttal to Strausz Et Al. Regarding Time-Resolved Fluorescence Depolarization of Asphaltenes*. 2009. p. 2845-2854.

57. Strausz, O. P.;Mojelsky, T. W., and Lown, E. M., *The Molecular Structure of Asphaltene: An Unfolding Story.* Fuel, 1992. **71**(12): p. 1355-1363.

58. Dickie, J. P. and Yen, T. F., *Macrostructures of the Asphaltic Fractions by Various Instrumental Methods.* Analytical Chemistry, 1967. **39**(14): p. 1847-1852.

59. Gray, M. R., *Consistency of Asphaltene Chemical Structures with Pyrolysis and Coking Behavior.* Energy & Fuels, 2003. **17**(6): p. 1566-1569.

60. Peng, P. a.;Morales-Izquierdo, A.;Hogg, A., and Strausz, O. P.,*Molecular Structure of Athabasca Asphaltene: Sulfide, Ether, and Ester Linkages*. 1997. p. 1171-1187.

61. Jaffe, S. B.;Freund, H., and Olmstead, W. N., *Extension of Structure-Oriented Lumping to Vacuum Residua.* Industrial & Engineering Chemistry Research, 2005. **44**(26): p. 9840-9852.

62. Boduszynski, M. M., *Composition of Heavy Petroleums. 2. Molecular Characterization.* Energy & Fuels, 1988. **2**(5): p. 597-613.

63. Boduszynski, M. M., *Limitations of the Average Structure Determination for Heavy Ends in Fossil Fuels.* Liquid Fuels Technology, 1984. **2**(3): p. 211-232.

64. Netzel, D. A., *NMR for Liquid Fossil Fuels -- Analytical Spectroscopy Library, Volume 1: Leonidas Petrakis and David Allen Elsevier, Amsterdam, 1987, Pp. 242, Isbn 0-444-42694-9, Dfl 175.00.* Fuel, 1987. **66**(9): p. 1310-1311.

65. Campbell, D. M. and Klein, M. T., *Construction of a Molecular Representation of a Complex Feedstock by Monte Carlo and Quadrature Methods.* Applied Catalysis A: General, 1997. **160**(1): p. 41-54.

66. Quann, R. J. and Jaffe, S. B., *Structure-Oriented Lumping: Describing the Chemistry of Complex Hydrocarbon Mixtures.* Industrial & Engineering Chemistry Research, 1992. **31**(11): p. 2483-2497.

67. Neurock, M.;Nigam, A.;Trauth, D., and Klein, M. T., *Molecular Representation of Complex Hydrocarbon Feedstocks through Efficient Characterization and Stochastic Algorithms.* Chemical Engineering Science, 1994. **49**(24, Part 1): p. 4153-4177.

68. Coates, J.,John Wiley & Sons Ltd, *Interpretation of Infrared Spectra, a Practical Approach*, in *Encyclopedia of Analytical Chemistry*, R.A. Meyers, Editor. 2000. p. 10815-10837.

69. Erich, K., *Far-Infrared Fourier Spectroscopy as a Method for Structure Determination in Chemistry.* Angewandte Chemie International Edition in English, 1976. **15**(1): p. 25-39.

70. Ramaswamy, V. and Singh, I. D., *Determination of N-Paraffinic Structure in Petroleum Fraction by Infrared Spectroscopy.* Fresenius' Journal of Analytical Chemistry, 1987. **327**(8): p. 774-776.

71. Pirali, O.;Van-Oanh, N.-T.;Parneix, P.;Vervloet, M., and Brechignac, P., *Far-Infrared Spectroscopy of Small Polycyclic Aromatic Hydrocarbons.* Physical Chemistry Chemical Physics, 2006. **8**(32): p. 3707-3714.

72. Zhang, K.;Guo, B.;Colarusso, P., and Bernath, P. F., *Far-Infrared Emission Spectra of Selected Gas-Phase Pahs: Spectroscopic Fingerprints.* Science, 1996. **274**(5287): p. 582-583.

73. A.L.Mattioda;Hudgins, D. M., and Allamandola, L. J., *Experimental near-Infrared Spectroscopy of Polycyclic Aromatic Hydrocarbons between 0.7 and 2.5 Micrometer.* The Astrophysical Journal, 2005. **629**: p. 1188-1210.

74. Yen, T. F.;Wu, W. H., and Chilingar, G. V., *A Study of the Structure of Petroleum Asphaltenes and Related Substances by Infrared Spectroscopy.* Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 1984. **7**(3): p. 203 - 235.

75. Wilt, B. K.;Welch, W. T., and Rankin, J. G.,*Determination of Asphaltenes in Petroleum Crude Oils by Fourier Transform Infrared Spectroscopy*. 1998. p. 1008-1012.

76. Blanco, M.;Maspoch, S.;Villarroya, I.;Peralta, X.;González, J. M., and Torres, J., *Determination of Physico-Chemical Parameters for Bitumens Using near Infrared Spectroscopy.* Analytica Chimica Acta, 2001. **434**(1): p. 133-141.

77. Hannisdal, A.;Hemmingsen, P. V., and Sjoblom, J., *Group-Type Analysis of Heavy Crude Oils Using Vibrational Spectroscopy in Combination with Multivariate Analysis.* Industrial & Engineering Chemistry Research, 2005. **44**(5): p. 1349-1357.

78. Aske, N.;Kallevik, H., and Sjoblom, J., *Determination of Saturate, Aromatic, Resin, and Asphaltenic (Sara) Components in Crude Oils by Means of Infrared and near-Infrared Spectroscopy.* Energy & Fuels, 2001. **15**(5): p. 1304-1312.

79. Scott, A. P. and Radom, L., *Harmonic Vibrational Frequencies: An Evaluation of Hartree-Fock, Moller - Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors.* The Journal of Physical Chemistry, 1996. **100**(41): p. 16502-16513.

80. Gaussian 03 Revision E.01, F., M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A.; Gaussian, Inc., Wallingford CT, 2004.

81. Foresman, J. B. and Frisch., A., *Exploring Chemistry with Electronic Structure Methods, 2nd Ed.* 1996, Pittsburgh, PA: Gaussian Inc.

82. Andersson, M. P. and Uvdal, P., *New Scale Factors for Harmonic Vibrational Frequencies Using the B3LYP Density Functional Method with the Triple-Î¶ Basis Set 6-311+G(D,P).* The Journal of Physical Chemistry A, 2005. **109**(12): p. 2937-2941.

83. Rauhut, G. and Pulay, P., *Transferable Scaling Factors for Density Functional Derived Vibrational Force Fields.* The Journal of Physical Chemistry, 2002. **99**(10): p. 3093-3100.

84. Liu, R.;Tate, D. R.;Clark, J. A.;Moody, P. R.;Van Buren, A. S., and Krauser, J. A., *Density Functional Theory Study of Molecular Structures and Vibrational Spectra of 3,4- and 2,3-Pyridyne.* The Journal of Physical Chemistry, 1996. **100**(9): p. 3430-3434.

85. El-Azhary, A. A. and Suter, H. U., *Comparison between Optimized Geometries and Vibrational Frequencies Calculated by the DFT Methods.* The Journal of Physical Chemistry, 1996. **100**(37): p. 15056-15063.

86. Kreibich, T.;van Leeuwen, R., and Gross, E. K. U., *Multicomponent Density-Functional Theory for Electrons and Nuclei.* Physical Review A (Atomic, Molecular, and Optical Physics), 2008. **78**(2): p. 022501-22.

87. Kohn, W., *Fundamentals of Density Functional Theory*, in *Density Functionals: Theory and Applications*. 1998. p. 1-7.

88. Szabo, A. and Ostlund, N. S., *Modern Chemistry: Introduction to Advanced Electronic Structure Theory*. 1982, New York: Macmillan Publishing C., Inc.

89. Cramer, C. J., *Essentials of Computational Chemistry; Theories and Models*. 2002, West Sussex: John Wiley & Sons Ltd.

90. Bachrach, S. M., *Computational Organic Chemistry*. 2007, New Jersey: John Wiley & Sons, Inc.

91. McQuarrie, D. A., *Quantum Chemistry*. 2nd ed. 2008, Sausalito, California: University Science Books.

92. Sherrill, C. D., (2000) *An Introduction to Hartree-Fock Molecular Orbital Theory (Http://Vergil.Chemistry.Gatech.Edu/Notes/Hf-Intro/Hf-Intro.Pdf); Accessed July 17, 2009*.

93. James, A. J., Imperial College of Science and Technology, *Solving the Many Electron Problem with Quantum Monte-Carlo Methods Ph.D. Thesis*, in *Condensed Matter Theory*. 1995: London.

94. Hohenberg, P. and Kohn, W., *Inhomogeneous Electron Gas.* Physical Review, 1964. **136**(3B): p. B864.

95. Kohn, W.;Becke, A. D., and Parr, R. G., *Density Functional Theory of Electronic Structure.* The Journal of Physical Chemistry, 1996. **100**(31): p. 12974-12980.

96. Kohn, W. and Sham, L. J., *Self-Consistent Equations Including Exchange and Correlation Effects.* Physical Review, 1965. **140**(4A): p. A1133.

97. John, A. P.;Peter, M. W. G., and Nicholas, C. H., *Spin-Unrestricted Character of Kohn-Sham Orbitals for Open-Shell Systems.* International Journal of Quantum Chemistry, 1995. **56**(4): p. 303-305.

98. Perdew, J. P.;Ruzsinszky, A.;Constantin, L. A.;Sun, J., and Csonka, G. b. I., *Some Fundamental Issues in Ground-State Density Functional Theory: A Guide for the Perplexed.* Journal of Chemical Theory and Computation, 2009. **5**(4): p. 902-908.

99. Kowalewski, I.;Vandenbroucke, M.;Huc, A. Y.;Taylor, M. J., and Faulon, J. L., *Preliminary Results on Molecular Modeling of Asphaltenes Using Structure Elucidation Programs in Conjunction with Molecular Simulation Programs.* Energy & Fuels, 1996. **10**(1): p. 97-107.

100. Jensen, F., *Introduction to Computational Chemistry*. 2nd ed. 2007, West Sussex: John Wiley & Sons.

101. Patricia, G.;Delia, S.-C., and Jesús, R.-O., *Performance of Dft Hybrid Functionals in the Theoretical Treatment of H-Bonds: Analysis Term-by-Term.* International Journal of Quantum Chemistry, 2008. **108**(2): p. 229-237.

102. Becke, A. D., *Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior.* Physical Review A, 1988. **38**(6): p. 3098.

103. Becke, A. D., *Density-Functional Thermochemistry. Iii. The Role of Exact Exchange.* The Journal of Chemical Physics, 1993. **98**(7): p. 5648-5652.

104. Lee, C.;Yang, W., and Parr, R. G., *Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density.* Physical Review B, 1988. **37**(2): p. 785.

105. Becke, A. D., *Density-Functional Thermochemistry. V. Systematic Optimization of Exchange-Correlation Functionals.* The Journal of Chemical Physics, 1997. **107**(20): p. 8554-8560.

106. Stephens, P. J.;Devlin, F. J.;Chabalowski, C. F., and Frisch, M. J., *Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields.* The Journal of Physical Chemistry, 2002. **98**(45): p. 11623-11627.

107. Perdew, J. P.;Chevary, J. A.;Vosko, S. H.;Jackson, K. A.;Pederson, M. R.;Singh, D. J., and Fiolhais, C., *Atoms, Molecules, Solids, and Surfaces: Applications of the Generalized Gradient Approximation for Exchange and Correlation.* Physical Review B, 1992. **46**(11): p. 6671.

108. Michaelian, K. H.;Billinghurst, B. E.;Shaw, J. M., and Lastovka, V., *Far-Infrared Photoacoustic Spectra of Tetracene, Pentacene, Perylene and Pyrene.* Vibrational Spectroscopy, 2009. **49**(1): p. 28-31.

109. Lastovka, V. and Shaw, J. M., *Predictive Correlation for Cp of Organic Solids Based on Elemental Composition.* Journal of Chemical & Engineering Data, 2007. **52**(4): p. 1160-1164.

110. Montgomery, D. C. and Runger, G. C., *Applied Statistics and Probability for Engineers 3rd Ed.* 2003: John Wiley & Sons, Inc.

Appendix I

Table A 1 -1: Table depicting the algorithm used to perform least squares optimization of frequency data generated

	LRC_1	⋯	LRC_n		CAA	
Freq.					*Measured*	
(cm^{-1})	AA_1	⋯	AA_n	*Predicted(P)*	*(M)*	$(M-P)^2$
f_1	IR_{11}	⋯	IR_{n1}	$\Sigma((LRC_1*IR_{11})+\cdots+(LRC_n*IR_{n1}))$	IR_1	$(M-P)_1^2$
f_2	IR_{12}	⋯	IR_{n2}	$\Sigma((LRC_1*IR_{12})+\cdots+(LRC_n*IR_{n2}))$	IR_2	$(M-P)_2^2$
⋮	⋮		⋮	⋮	⋮	⋮
f_i	IR_{1i}	⋯	IR_{ni}	$\Sigma((LRC_1*IR_{1i})+\cdots+(LRC_n*IR_{ni}))$	IR_i	$(M-P)_i^2$
					Objective function =	$\Sigma((M-P)_1^2+(M-P)_2^2+\cdots+(M-P)_i^2)$

LRC = Linear Regression Coefficient for the *n* pure aromatic ring or aliphatic chain compound in library (can have only positive values)
CAA = Combination of aromatic ring and aliphatic chain compounds used to model ill-defined hydrocarbon structures (independent variable) e.g. *pyrene + decane + pyrene*
AA = Pure aromatic ring or aliphatic chain compounds in the library (dependent variable) e.g. pyrene, decane
f_i = Computed (or measured) frequencies at $10cm^{-1}$ intervals,
IR_i = Computed infrared intensity or Raman activity for a given frequency
Predicted (P) = Arithmetic sum of the product of regression coefficients and infrared intensities or Raman activities for each frequency, *i*
Measured (M) = Computed infrared intensity or Raman activity of the model ill-defined hydrocarbon compound for a given frequency
Objective function = Sum of squares when $LRC \neq 0$

Calculating the Residuals:

$$\text{Residual} = \frac{\sum (IR \text{ or } Raman \text{ of } CAA)^2}{\sum (M - P)^2} \quad \text{(A1.1)}$$

where,

$$\sum (IR \text{ or } Raman \text{ of } CAA)^2 \Rightarrow LRC = 0 \quad \text{(A1.2)}$$

Table A 1 -2: Table showing the results of the analysis of residuals for the 20 model compounds based on infrared spectra

	Sum of squares		
Model Compound	*LRC*=0	*LRC*≠ 0	Residual
anthracene + decane + chrysene	*2.03E+06*	*8.85E+05*	*44%*
anthracene + decane + pyrene	*2.24E+05*	*4.72E+04*	*21%*
anthracene + octane + pyrene	*1.74E+05*	*4.39E+04*	*25%*
anthracene + dodecane + pyrene	*2.91E+05*	*6.15E+04*	*21%*
chrysene + decane + fluoranthene	*2.72E+05*	*1.04E+05*	*38%*
chrysene + decane + phenanthrene	*2.39E+05*	*9.55E+04*	*40%*
chrysene + octane + pyrene	*1.89E+05*	*5.10E+04*	*27%*
chrysene + dodecane + pyrene	*3.17E+05*	*7.15E+04*	*23%*
fluoranthene + decane + phenanthrene	*2.63E+05*	*9.93E+04*	*38%*
pyrene + decane +pyrene	*2.63E+05*	*4.89E+04*	*19%*
pyrene + dodecane +pyrene	*3.19E+05*	*6.03E+04*	*19%*
pyrene + octane +pyrene	*1.89E+05*	*3.21E+04*	*17%*
naphthalene + decane + fluoranthene	*2.49E+05*	*9.50E+04*	*38%*
naphthalene + decane + pyrene	*1.99E+05*	*3.50E+04*	*18%*
naphthalene + decane + phenanthrene	*2.17E+05*	*8.47E+04*	*39%*
naphthalene + octane + pyrene	*1.48E+05*	*2.32E+04*	*16%*
naphthalene + dodecane + pyrene	*2.71E+05*	*4.80E+04*	*18%*
phenanthrene + decane + pyrene	*2.17E+05*	*4.00E+04*	*18%*
phenanthrene + octane + pyrene	*1.75E+05*	*3.86E+04*	*22%*
phenanthrene + dodecane + pyrene	*2.85E+05*	*6.05E+04*	*21%*

Table A 1 -3: Table showing the results of the analysis of residuals for the 20 model compounds based on Raman spectra

	Sum of squares		
Model Compound	*LRC*=0	*LRC*≠ 0	Residual
anthracene + decane + chrysene	*2.23E+04*	*8.13E+03*	*36%*
anthracene + decane + pyrene	*5.25E+06*	*9.60E+05*	*18%*
anthracene + octane + pyrene	*5.08E+06*	*1.05E+06*	*21%*
anthracene + dodecane + pyrene	*5.57E+06*	*9.99E+05*	*18%*
chrysene + decane + fluoranthene	*8.91E+06*	*1.06E+06*	*12%*
chrysene + decane + phenanthrene	*8.12E+06*	*1.21E+06*	*15%*
chrysene + octane + pyrene	*9.22E+06*	*1.67E+06*	*18%*
chrysene + dodecane + pyrene	*9.82E+06*	*1.74E+06*	*18%*
fluoranthene + decane + phenanthrene	*5.17E+06*	*1.01E+06*	*19%*
pyrene + decane + pyrene	*7.14E+06*	*1.57E+06*	*22%*
pyrene + dodecane +pyrene	*7.60E+06*	*1.61E+06*	*21%*
pyrene + octane +pyrene	*6.89E+06*	*1.53E+06*	*22%*
naphthalene + decane + fluoranthene	*3.88E+06*	*3.64E+05*	*9%*
naphthalene + decane + pyrene	*3.40E+06*	*4.18E+05*	*12%*
naphthalene + decane + phenanthrene	*2.86E+06*	*4.90E+05*	*17%*
naphthalene + octane + pyrene	*3.17E+06*	*4.28E+05*	*14%*
naphthalene + dodecane + pyrene	*3.73E+06*	*4.62E+05*	*12%*
phenanthrene + decane + pyrene	*4.22E+06*	*4.82E+05*	*11%*
phenanthrene + octane + pyrene	*4.01E+06*	*5.02E+05*	*13%*
phenanthrene + dodecane + pyrene	*4.54E+06*	*5.25E+05*	*12%*

Appendix II

Appendix II consists of the files used to perform the regression analysis of all the model compounds based on infrared and Raman spectra (to view these files, please contact: jmshaw@ualberta.ca).

Appendix III

Appendix III consists of the files used to perform the regression analysis of all the model compounds based on proton and carbon-13 NMR spectra (to view these files, please contact: jmshaw@ualberta.ca).