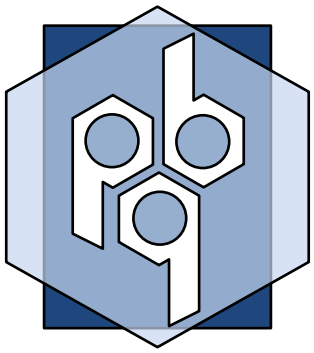


Engenharia de Processos Químicos e Bioquímicos
Escola de Química - UFRJ

MOLECULAR RECONSTRUCTION OF HEAVY PETROLEUM FRACTIONS

Helton Siqueira Maciel

Master's Thesis presented to Engenharia de Processos Químicos e Bioquímicos Graduate Program, Escola de Química, Universidade Federal do Rio de Janeiro, as a partial fulfillment of the requirements for the degree of Master of Science

Advisors: Frederico Wanderley Tavares
Charlles Rubber de Almeida Abreu

Rio de Janeiro
September 2019

MOLECULAR RECONSTRUCTION OF HEAVY PETROLEUM FRACTIONS

Helton Siqueira Maciel

THESIS SUBMITTED TO THE FACULTY OF ENGENHARIA DE PROCESSOS QUÍMICOS E BIOQUÍMICOS GRADUATE PROGRAM OF UNIVERSIDADE FEDERAL DO RIO DE JANEIRO AS A PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN CHEMICAL ENGINEERING.

Examined by:

Prof. Frederico Wanderley Tavares, D.Sc.

Prof. Charlles Rubber de Almeida Abreu, D.Sc.

Prof. Argimiro Resende Secchi, D.Sc.

Prof. Rafael de Pelegrini Soares, D.Sc.

RIO DE JANEIRO, RJ – BRAZIL
SEPTEMBER 2019

Maciel, Helton Siqueira

Molecular reconstruction of heavy petroleum fractions/Helton Siqueira Maciel. – Rio de Janeiro: UFRJ/EQ, 2019.

XIX, 116 p.: il.; 29, 7cm.

Advisors: Frederico Wanderley Tavares

Charlles Rubber de Almeida Abreu

Dissertation (master) – UFRJ/EQ/Engenharia de Processos Químicos e Bioquímicos Graduate Program, 2019.

Bibliography: p. 107 – 116.

1. Stochastic Reconstruction. 2. Molecular Characterization. 3. Simulator-Based Models. 4. Approximate Bayesian Computation. I. Tavares, Frederico Wanderley *et al.* II. Universidade Federal do Rio de Janeiro, Escola de Química, Engenharia de Processos Químicos e Bioquímicos Graduate Program. III. Title.

"Instead of fearing wrong predictions, we look eagerly for them; it is only when predictions based on our present knowledge fail that probability theory leads us to fundamental new knowledge."

— E. T. Jaynes

Aknowledgements

I would like to thank my parents, Ricardo (*in memoriam*) and Maristela for the love, support and education and my brother Hudson for the partnership. I also would like to thank my wife, Jéssica, for supporting this endeavour. To my advisors, Charlles and Fred, I would like to thank for the discussions and guidance in developing this work. I also would like to thank CENPES bottom of the barrel research group , Adriana, Danielle, Diego and Natalie, for the discussions and the support, specially Diego, who worked together with me in the development of the algorithms. Finally, i would like to thank PETROBRAS for allowing me to conclude this work.

Resumo da Dissertao apresentada à EPQB/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

RECONSTRUÇÃO MOLECULAR DE FRAÇÕES PESADAS DE PETRÓLEO

Helton Siqueira Maciel

Setembro/2019

Orientadores: Frederico Wanderley Tavares
Charlles Rubber de Almeida Abreu

Programa: Engenharia Qumica

Apresenta-se, neste trabalho, uma metodologia para construir moléculas de frações pesadas de petróleo com base em dados experimentais limitados. O algoritmo desenvolvido é dividido em três etapas. **1)** O processo de geração molecular é baseado no algoritmo de reconstrução estocástica. A estimação de parâmetros de modelos de reconstrução estocástica é um desafio devido às suas verossimilhanças intratáveis. A inferência de parâmetros foi tratada a partir de uma perspectiva Bayesiana usando a estrutura de otimização bayesiana para inferência livre de verossimilhança. **2)** Uma técnica de agrupamento não hierárquico foi desenvolvida para escolher um subconjunto de moléculas representativas do conjunto molecular inicial gerado a partir do algoritmo de reconstrução estocástica. **3)** Para o cálculo da composição, foi aplicada a reconstrução pelo método de maximização de entropia. Aplicamos a nossa metologia a diferentes resíduos de vácuo de diferentes origens. O modelo foi capaz de representar os resíduos de vácuo estudados neste trabalho. Além de replicar os dados a partir dos quais foi treinado, o modelo também foi capaz de prever efetivamente novas propriedades dessas misturas complexas.

Abstract of Thesis presented to EPQB/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

MOLECULAR RECONSTRUCTION OF HEAVY PETROLEUM FRACTIONS

Helton Siqueira Maciel

September/2019

Advisors: Frederico Wanderley Tavares
Charlles Rubber de Almeida Abreu

Department: Chemical Engineering

In this work, we present a methodology to build molecules of heavy petroleum fractions based on limited experimental data. Our algorithm is divided into three steps. **1)** The molecular generation process is based on the stochastic reconstruction algorithm. The parameter estimation of stochastic reconstruction models is a challenge due to their intractable likelihoods. We evaluated the parameter inference from a Bayesian perspective using the Bayesian optimization framework for likelihood-free inference. **2)** We developed a non-hierarchical clustering technique to choose a subset of representative molecules from the initial molecular ensemble generated from the stochastic reconstruction algorithm. **3)** For composition calculation, we applied the reconstruction by entropy maximization method. We applied our methodology to different vacuum residues from different origins. The model was able to represent the vacuum residues studied in this work. Besides replicating the data from which it was trained, the model was also able to effectively predict new properties of these complex mixtures.

Contents

List of Figures

List of Tables

List of Symbols

$p(x)$	Probability density function for a generic random variable x.
$P(x)$	Cumulative density function for a generic random variable x.
χ^2	A random variable that follows a Chi-squared probability density function.
Γ	Gamma function.
ν	Degrees of freedom for the Chi-squared probability density function.
β	Shape parameter for the exponential probability density function.
$\mathbf{M}$	Molecular matrix using the structure-oriented lumping representation.
$\mathbf{S}$	Stoichiometry matrix for the structure-oriented lumping representation.
$\mathbf{MA}$	Matrix representing the quantity of each atom type for each molecule.
$\mathbf{M_{ex}}$	Extended molecular matrix using the Structure-oriented lumping representation.
$\mathbf{F}$	Matrix representing the relationship between functional groups and chemical attributes.
$\mathbf{MF}$	Matrix representing the quantity of each functional group for each molecule.
$\mathbf{VMW}$	Vector of molecular masses for each atom type.
$\mathbf{MMW}$	Vector of molecular masses for each molecule.
T_b	Normal boiling point for each molecule.
F_{T_b}	Correction factor for the normal boiling point calculation.
N_R	Number of rings for each molecule.
d	Specific gravity for each molecule.

F_{V_m}	Correction factor for the specific gravity calculation.
MPM	Matrix containing each molecule structural contribution to the boiling point and specific gravity.
FGP	Matrix containing each functional group contribution to the boiling point and specific gravity.
x_i	Molar fraction.
MW_{avg}	Average molecular mass.
$\mathbf{X}$	Molar fraction vector.
w_i	Mass fraction.
$S_{g_{mix}}$	Mixture specific gravity.
w_i	Volume fraction.
cw_i	Cumulative mass fraction.
cv_i	Cumulative volume fraction.
$T_{b,k\%}$	Boiling point associated with a k % vaporization.
$\%H_i$	Molecule hydrogen mass percentage.
$w_{k_{mix}}$	Mass fraction of a atom type k in the mixture.
$x_{k_{mix}}$	Molar fraction of a atom type k in the mixture.
$L(\theta\|y)$	Likelihood function.
$\pi(\theta\|y)$	Posterior distribution.
$\pi(\theta)$	Prior distribution.
$d(y_\theta, y_0)$	Discrepancy between observed and simulated data.
ϵ	Discrepancy threshold.
$L_{d,\epsilon}(\theta)_{abc}$	Approximate likelihood function.
$\mathcal{GP}$	Gaussian process prior function.
$k(\theta, \theta')$	Squared exponential covariance function.
$\mathcal{E}_{1:t}$	Evidence set used for the Gaussian process training.
$m_{1:t}$	Posterior mean of the Gaussian process given evidence $\mathcal{E}_{1:t}$.
$\nu^2_{1:t}(\theta)$	Posterior variance of the Gaussian process given evidence $\mathcal{E}_{1:t}$.
$\pi^{ABC}(\theta\|\mathcal{E}_{1:t})$	Approximate posterior distribution.
$\mathbb{E}(\pi^{ABC}(\theta\|\mathcal{E}_{1:t}))$	Expected value of the unnormalized approximate posterior distribution.
$\mathbb{V}(\pi^{ABC}(\theta\|\mathcal{E}_{1:t}))$	Variance of the unnormalized approximate posterior distribution.
$T(h, a)$	Owen's t-function.

$\hat{R}$	Potential scale reduction for Markov chain simulations.
$\hat{n}_{eff}$	Number of effective samples for Markov chain simulations.
θ	PDF's parameters.
φ	Rescaled PDF's parameters.
E	Entropy of a generic probability distribution.
H	Entropy of a molar fraction based probability distribution.
λ	Lagrangian multipliers.
$\mathbf{g}$	Constraints matrix for the reconstruction by entropy maximization method.
c	Number of clusters for the partitioning around medoids method.
$d_{i,i'}$	Euclidean distance to measure dissimilarity between molecule i and molecule i'.
$u_{i'}$	Normalized total distance between molecule i' and the rest of the molecules.
$p(\tilde{y}\|y_0)$	Posterior predictive distribution.

List of Algorithms

Chapter 1

Introduction

Petroleum and its derivatives still play a central role in the energy sector. The advances of alternative energy sources in the past years motivated the oil industry to be more efficient in its operations (VASSILIOU, 2018). Petroleum refining is a central part of the oil chain of production. Most of the products used by the energy sector are produced in refineries. A refinery is a complex industrial plant with many process units (COKER, 2018).

The main goal of a refinery is to transform crude oil into more valuable products. That can be done by physical separation processes, such as distillation and solvent extraction, or by chemical conversion processes, either thermal or catalytic (SPEIGHT and OZUM, 2001). The profitability of a refinery is intimately connected with its conversion capacity. In that sense, units responsible for the conversion of the heavier fractions of petroleum, such as the fluid catalytic cracking and delayed coking units, play a central role in the refinery margin (GRAY, 2003).

Optimal design and operation of such conversion units may greatly increase refinery margin and profitability. For that, modeling and simulation of refining processes is a powerful tool. More than that, representative models can guide the whole refinery or even a whole market to a better place in terms of oil allocation, process operation and products marketing and distribution (COKER; GRAY; SPEIGHT; SPEIGHT and OZUM; VASSILIOU, 2018; 1994; 2014; 2001; 2018).

To accurately model a refining unit, one should be able to characterize the petroleum fractions involved in that operation. Due to its chemical complexity, the characterization of petroleum fractions has been a challenge since the early days of the industry. The first approach in this matter was to use oil physical properties, such as boiling point, molecular weight, specific gravity and solubility, and then translate that information in terms of pseudo components (RIAZI, 2005).

Although the pseudo component approach was very successful in the representation of separation processes, its application to the conversion processes showed severe limitations. The lack of molecular detail restricted the use of such models to

the data used in its development. Besides that, any major change in the process, for instance, the catalyst type, required a complete reparametrization of the model (ANCHEYTA *et al.*; DE OLIVEIRA, LUÍS P. *et al.*; DENIZ *et al.*; WEI *et al.*, 2005; 2016; 2017b; 2008).

To overcome those limitations, a molecular-based approach in the development of petroleum conversion processes models is necessary. However, despite the advances in the field of analytical chemistry, only a broad perspective of the chemical structure of oil fractions is attainable (MCKENNA *et al.*; MCKENNA *et al.*; MCKENNA *et al.*; MCKENNA *et al.*; PODGORSKI *et al.*, 2010a; 2010b; 2013a; 2013b; 2013). For that, methods designed to mimic the molecular composition of petroleum fractions based on general experimental information and prior chemical knowledge have been developed. We shall refer to these methods as molecular reconstruction techniques.

In a first attempt, QUANN and JAFFE (1992) proposed a method called structure-oriented lumping. This technique consisted of a vector representation of petroleum molecules using a predefined set of molecular attributes, such as the number of benzenes or cyclopentane. TRAUTH *et al.* (1994) proposed a model based on the representation of chemical attributes by probability density functions. Those probability density functions can then be sampled by a Monte Carlo procedure. When coupled with an optimization loop for the parameters of the distributions, the stochastic reconstruction method arises. Inspired by the structure-oriented lumping method, PENG (1999) developed a molecular reconstruction technique called molecular type homologous series. HUDEBINE *et al.* (2002) included a second step in the stochastic reconstruction algorithm called reconstruction by entropy maximization.

In this work, we developed a novel molecular reconstruction algorithm to be applied to heavy petroleum fractions. Our method combines the structure-oriented lumping, stochastic reconstruction and reconstruction by entropy maximization methods. Furthermore, we included a third step in the algorithm, in which we use a non-hierarchical clustering technique to choose the best molecular candidates from the entire molecular ensemble.

The chapters of this thesis are divided as follows: In Chapter 2, we present the objectives of this work. In Chapter 3 a literature review of the relevant work for the scope of this thesis. In Chapter 4, we describe the molecular reconstruction algorithm developed here. In Chapter 5, we describe the parameter inference procedure. In chapter 6, we present the results of the application of the algorithm to the reconstruction of different vacuum residues. At last, in chapter 7, we give our final remarks on the work.

Chapter 2

Objectives

The main goal of this work is to develop a methodology to mimic the molecular composition of heavy petroleum fractions based on general (and limited) experimental information.

- For the molecular generation, we developed a model based on the stochastic reconstruction algorithm (TRAUTH *et al.*, 1994). Such algorithms are based on the modeling of chemical attributes using probability density functions.
- We combined the flexibility of the stochastic reconstruction algorithm with the convenient framework of the structure-oriented lumping for molecular representation.
- We proposed an extension of the structure-oriented lumping vector to improve molecular diversity.
- We analyzed the parameter inference of the stochastic reconstruction model from a bayesian perspective.
- We developed a non-hierarchical clustering technique to select a subset of representative molecules from the initial molecular ensemble.
- We calculated the molecular composition using the reconstruction by entropy maximization method proposed by HUDEBINE *et al.* (2002).

Chapter 3

Literature Review

In this chapter, we discuss the relevant literature for the scope of this work. The chapter is divided into three topics. First, we pass through the analytical developments in the realm of heavy petroleum fractions characterization. The molecular reconstruction algorithms heavily rely on a general knowledge of petroleum chemistry, which serves as a base for the model of construction. Second, we review the works focused on the molecular reconstruction itself. Then, we talk about a major part of the reconstruction algorithms, the estimation of the model parameters. We finish this chapter outlining the contributions of this thesis to the literature.

3.1 Chemistry of Heavy Petroleum Fractions

The interest in developing analytical techniques to characterize petroleum fractions, especially the heavy ones, is due to its utility in the design and optimization of refining processes. For this work, these results give a broad perspective of the molecular families, structures and functional groups, which serve as prior knowledge in the model building process. The petroleum fractions, in its molecular level, are usually called hydrocarbons due to the predominant content of carbon and hydrogen atoms. However, these fractions also contain a small but relevant quantity of the so-called heteroatoms: sulfur, nitrogen, and oxygen which can play a major role in the performance of the refining processes.

Focusing on identifying the different chemical families in the heavy petroleum fractions, LUMPKIN (1956) proposed a method to identify saturated hydrocarbons in heavy fractions using the mass spectrometer. They divided the fraction into specific classes: paraffins, noncondensed naphthenes, and condensed naphthenes. MEAD (1968) used a field ionization mass spectrometer to analyze paraffin waxes in the boiling range of $300\,^{\circ}C$ to $550\,^{\circ}C$ identifying normal paraffins, isoparaffins and alkylbenzenes. Besides that, they were able to quantify the carbon number, which ranged from 20 to 40 carbons.

SAWATZKY *et al.* (1976) proposed a method for the separation of heavy petroleum hydrocarbons into structural types - saturates, monoaromatics, diaromatics, and polyaromatics. Since it goes beyond classification, it gives an insight into the relative quantities of these groups. TRESTIANU *et al.* (1985) described a method to perform a simulated distillation of heavy petroleum fractions up to $800\,^{\circ}C$. In its results, we can see the overall shape of the distribution of boiling points in these heavy fractions which in turn can be extended to the shape of the carbon number distribution.

Entering into the heteroatom characterization, ROSE and FRANCISCO (1987) proposed a method to identify acid heteroatoms in heavy petroleum fractions. They have tested two vacuum residua, a name given to the bottom product of the vacuum distillation unit usually with boiling point starting at $550\,^{\circ}C$, and the n-heptane insoluble fraction (asphaltenes) from one of these residua. Qualitatively, the main acidic functional groups identified were the hydroxyl (-OH), the carboxylic acid (-COOH), imino (=NH), and thiol (-SH). DUTRIEZ *et al.* (2010) measured the composition of heavy petroleum fraction in terms of molecular groups using two-dimensional gas chromatography. He divided the fractions into saturates, monoaromatics, diaromatics, triaromatics and tetraaromatics+. Besides composition, the boiling point distributions of such families are reported.

In an attempt of giving a more detailed molecular description of heavy petroleum fractions, Boduszynski and collaborators published a series of four papers. In the first paper, BODUSZYNSKI (1987) studied the variation of molecular weight, hydrogen deficiency, and heteroatom concentrations as functions of the atmospheric equivalent boiling point (AEBP). Besides sulfur, nitrogen, and oxygen, the author also considers the most abundant metals in the heteroatom classification, such as nickel, vanadium, and iron. The proposed methodology was applied to the atmospheric residue fraction, which is the bottom product of the atmospheric distillation unit with a boiling point starting at around $390\,^{\circ}C$. BODUSZYNSKI (1987) concludes that heavy petroleums, and residues in particular, are not composed mostly of very high molecular weight components. The results reveal that most heavy petroleum components do not exceed a molecular weight of approximately 2000. He also concludes that the heteroatom concentrations and hydrogen deficiency increase with increasing AEBP. Significant bimodal distribution patterns for nickel and vanadium were observed.

In the second work, BODUSZYNSKI (1988) tried to describe the chemical composition as a function of the atmospheric equivalent boiling point. The molecular types classification given by the author was heavily used here. He divides the heavy fractions into three major types: alkanes (paraffins), cycloalkanes (naphthenes), and aromatic hydrocarbons. Besides that, we have also considered the proposed nitro-

gen occurrence, mainly divided into basic nitrogen (pyridine) and pyrrolic nitrogen BODUSZYNSKI (1988). The remaining two papers, ALTGELT and BODUSZYNSKI (1992) and BODUSZYNSKI and ALTGELT (1992), addressed a boiling point-molecular weight correlation for distillable and non-distillable heavy fractions, respectively, where the authors propose that crude oil is a continuum in molecular weight, structure, and boiling point, even though they could not fully support this hypothesis from experimental results.

ROUSSIS and PROULX (2002) obtained the molecular weight distribution for heavy petroleum fractions using different methodologies. In the opposite direction of what was suggested in BODUSZYNSKI (1987), molecules with molecular weights up to 7000 were observed. In a different study, ROUSSIS and PROULX (2004) measured the molecular weight of non-boiling petroleum fractions detecting molecules with molecular weights up to 20000. However, this time the authors attribute these high molecular weights structures to an aggregation phenomenon, since the abundance of such molecules reduces in experimental conditions that are favorable to dissociation. QIAN *et al.* (2007) proposed different experimental methodologies to measure the molecular weight of heavy petroleum fractions. They detected molecules with molecular weight up to 5000, however, they also attribute these numbers to molecular aggregation.

MCKENNA *et al.* (2010b) developed an experimental methodology to support the Boduszynski model, confirming its validity to the heavy vacuum gas oil cut, considered a middle distillate. In the second paper, MCKENNA *et al.* (2010a) extended the experimental analysis to temperatures beyond the middle distillate cut. However, as stated by the authors, projection of distillable compositional space to higher carbon number cannot accurately describe non- distillable due to incompatibility with bulk asphaltene H:C ratios. The inescapable conclusion is that either asphaltene (non-distillable) are not high molecular weight species, or the continuity model does not apply to nondistillable materials. In the third paper, MCKENNA *et al.* (2013a) discuss the asphaltenes aggregation, observing that most asphaltenes are non-covalently aggregated. In the fourth paper, MCKENNA *et al.* (2013b) did a more detailed evaluation of the asphaltenes compositional space, concluding that asphaltenes (non-distillable) are not an extension of the distillable compositional space to higher and higher carbon number but an extension to higher degrees of aromaticity. Regarding the molecular weight of the heavier fractions, the results indicate that values do not exceed 2000, in agreement with BODUSZYNSKI (1987).

For the identification of acids in heavy petroleum fractions, QIAN *et al.* (2001a) proposed an experimental methodology based on mass spectrometry. Experimental results show the main functional groups present in these acidic structures. In general, they have the presence of oxygen and sulfur atoms in the form of carboxylic acids

and thiophene structures, respectively. Continuing his work, QIAN *et al.* (2001b) studied the nitrogen-containing aromatic compounds in heavy petroleum fractions. In agreement with the work of BODUSZYNSKI (1988), the main forms of nitrogen occurrence are the basic nitrogen (pyridine) and pyrrolic nitrogen. Also studying the speciation of nitrogen compounds in heavy cuts, DUTRIEZ *et al.* (2011) proposed a methodology base on a two-dimensional gas chromatography. The work supports main nitrogen occurrence classes: pyridinic and pyrrolic cores. Besides that, the author suggests that the nitrogen-containing compounds in heavy petroleum cuts are usually composed of highly alkylated polyaromatics structures, such as carbazoles, benzocarbazoles, dibenzocarbazoles (neutrals) and acridines, benzoacridines or dibenzoacridines (basics).

Another approach one can take when trying to better characterize petroleum fractions, including the heavy ones, is to study its reactivity. Since the mechanisms are built from verified elementary steps, one can determine the general form of the structure of the reactants based on product distribution. GRAY and MCCAFFREY (2002) studied the chain reactions and olefin formation in cracking, hydroconversion, and coking of petroleum and bitumen fractions. According to the author, the residue fraction contains more than 60 % of the carbon in saturated chain and ring structures. The author also states that as much as 40 % of the sulfur present occurs as reactive thioethers and thiolanes in saturated structures. Regarding the general molecular structure, the author states that an effective chemical model for asphaltenes and other components in the residue fraction is a random copolymer of aromatic cores joined by bridges and attached to pendant groups. GRAY (2003) discuss the consistency of asphaltene chemical structures with pyrolysis and coking behavior. Observing the nature of the products from mild and severe thermal cracking, the most consistent general form of asphaltenes are aromatic groups joined by bridges and substituted by aliphatic groups.

In this section, we reviewed the most relevant work, for this thesis, in terms of heavy petroleum chemistry. The analytical results give a broad perspective on the chemical families, functional groups, relative quantities, the general shape of properties distributions among other crucial information. This chemical knowledge should comprise the basic building blocks of any molecular reconstruction method, as it is the case of this thesis.

3.2 Molecular Reconstruction Methods

In this work, we define molecular reconstruction as a technique that tries to mimic the molecular composition of any petroleum fractions purely from general (bulk) experimental results and prior chemical knowledge. It can estimate both molecular

structures and composition. As shown in Section 3.1, although analytical procedures are capable of giving a general perspective on the petroleum chemistry, they are insufficient to fully characterize some fractions, especially the heavy ones. For that, molecular reconstruction methods play a major role in the development of molecular-level models for design, evaluation, and optimization of refining process, mainly the ones that involve chemical reactions.

One of the first methodologies for molecular-based modeling and molecular reconstruction was described in QUANN and JAFFE (1992). The authors proposed a method called Structure-oriented lumping (SOL). This technique represents individual hydrocarbon molecules as a vector of incremental structural features. In this manner, a mixture of hydrocarbons is represented as a set of these vectors. Structure-oriented lumping defines the basic building blocks of petroleum molecules. However, no definitive methodology on how to combine these blocks are given. One can see the methodology as a convenient framework for constructing molecular mixtures, calculate their properties and construct reaction networks. In QUANN and JAFFE (1996) and QUANN (1998), the authors explore the use of the Structure-oriented lumping framework to build molecular-based kinetic models. An extension of the structure-oriented lumping method was proposed in JAFFE *et al.* (2005). To better represent vacuum residues, the authors included metallic groups and a methodology to represent multi-core molecules.

Since then, many researchers have used the structure-oriented lumping to build kinetic models. CHRISTENSEN *et al.* (1999) used the structure-oriented lumping to build molecular models for a fluid catalytic cracking unit. The authors used more than 3000 molecules and over 60 reaction rules. YANG *et al.* (2008) used the structure-oriented lumping to simulate the secondary reactions of fluid catalytic cracking gasolines. TIAN *et al.* (2010) applied the methodology to build a steam cracking model. In two papers, TIAN *et al.* (2012a) and TIAN *et al.* (2012b) developed a delayed coking model based on the structure-oriented framework. Most researchers that use the SOL framework relies on prior chemical knowledge to build representative molecular cores for the fractions in question. They also rely on the concept of homologous series, a series of molecules of the same type with different carbon numbers, to build the complete mixture. Although very convenient, the structure-oriented lumping is limited when it comes to molecular diversity. The fixed molecular attributes proposed to wind up limiting the configuration and functional groups of the formed molecules. This can be problematic, especially for the heavier fractions.

Another popular methodology is the molecular type homologous series matrix (MTHS) described in PENG (1999). Different from the structure-oriented lumping, MTHS defines chemical cores not only attributes. Moreover, it proposes that the

petroleum mixtures are composed of homologous series of these chemical cores. In a sense, the MTHS method defines all the structures that could be present in the petroleum fractions, leaving the composition as a degree of freedom. The composition is often estimated through the definition of an objective function comparing calculated and experimental data. It is clear that it also suffers from a lack of molecular diversity.

HU *et al.* (2002) extended the MTHS application to refinery optimization, introducing the concept of molecular management of refining operations. AYE and ZHANG (2005) proposed a methodology based on the MTHS matrix. An automatic method to translate the physical properties of a hydrocarbon stream to the molecular information of the matrix is developed and successfully applied for gasoline-range fractions. GOMEZ-PRADO *et al.* (2008) proposed a modified MTHS matrix to represent any hydrocarbon stream. The fraction of each component in the stream is computed by minimizing the discrepancies between bulk and calculated characterization parameters. Furthermore, the authors propose a methodology to transform the information into a useful input for hydrocarbon lumped kinetic models. WU and ZHANG (2010) developed a methodology based on the MTHS matrix to represent gasoline and diesel streams. Besides the modification of the matrix itself, the authors considered that the composition and properties of molecular homologous series can be represented by probability distribution functions (PDF), changing the way to transform experimental information into a molecular composition. PYL *et al.* (2011) used the concept of homologous series of components to model crude oil fractions. The authors imposed probability density functions on both the carbon number distribution in each homologous series of components and on the structural attribute distributions. AHMAD *et al.* (2011) extended the use of the MTHS method to heavier petroleum fractions. The main difference from other works is the use of group contribution methods and mixing rules to calculate mixture properties.

The MTHS methodology poses an elegant way of representing molecules in petroleum fractions. However, it suffers from the same problems as the structure-oriented lumping technique, the lack of molecular diversity. Besides that, the original method requires a direct estimation of the molecular composition, falling into overfitting issues. This issue is partially solved by the use of probability density functions. Nevertheless, these limitations tend to be critical, especially for heavier fractions, the main subject of this thesis.

Another development in the field of molecular reconstruction methods was based on the representation of complex mixtures properties and possibly chemical structures with probability density functions. The use of probability density functions to represent molecular properties dates back to FLORY (1952), who showed that the molecular weight distributions of polymers could be modeled as a gamma distribu-

tion. The experimental development discussed in Section 3.1, specially the ones from Boduszynski and McKenna (ALTGELT and BODUSZYNSKI; BODUSZYNSKI; BODUSZYNSKI; BODUSZYNSKI and ALTGELT; MCKENNA *et al.*; MCKENNA *et al.*; MCKENNA *et al.*; MCKENNA *et al.*; PODGORSKI *et al.*, 1992; 1987; 1988; 1992; 2010a; 2010b; 2013a; 2013b; 2013), also provides experimental support for the use of probability density functions to represent molecular weight, boiling point, and carbon number distribution of petroleum fractions .

NEUROCK *et al.* (1990) extended the concept of probability density functions from properties to chemical attributes, originating a class of molecular reconstruction methods known as stochastic reconstruction. By chemical attributes we mean the basic building blocks of a molecule, such as number of cores, number of aromatic rings, ring configuration and so forth. Logically combining these probability distributions, one can sample from their cumulative form using a Monte Carlo type procedure. The Monte Carlo sampler is responsible for the stochastic nature of the technique. NEUROCK *et al.* (1990) proposed a series of chemical attributes probability distribution functions for the reconstruction of the n-heptane insoluble fraction (asphaltenes). The distributions are as follows: Number of cores, number of aromatic rings, the configuration of aromatic cores, naphthenic rings, degree of aliphatic substitution, length of aliphatic side chain, from which they sampled 10000+ molecules. However, the authors did not define the distributions shape, parameters or how one can estimate that.

In NEUROCK *et al.* (1994), the authors describe the stochastic reconstruction algorithm as a four-step procedure. The deduction of a chemical logical diagram, the compilation of structural cumulative probability distributions, stochastic sampling of each distribution and the construction of the molecular species. Regarding the parameters of the probability density functions, the authors use the average molecular parameter concept together with correlations developed elsewhere (HIRSCH and ALTGELT, 1970). In some sense, at the time being, the method is restricted to a molecular generation procedure instead of a molecular reconstruction, since it did not estimate the PDF's parameters.

TRAUTH *et al.* (1994) coupled the Monte Carlo molecular generation with an optimization step to find the best values for the PDF's parameters. For that, a chi-square objective function comparing calculated and experimental data were used. The shape of most distribution functions are simplified by a chi-squared functional form, a particular case of the gamma distribution. In this procedure, only average (bulk) properties of the petroleum fractions are used in the objective function. The authors used seven attributes to construct a residue: paraffin length, number of naphthenic rings, number of side chains, side-chain length, number of aromatic rings, number of naphthenic rings per aromatic molecule and asphaltene degree

of polymerization. The stochastic reconstruction algorithm assumes an equimolar mixture from the molecules generated.

HUDEBINE *et al.* (2002) developed an algorithm to calculate the composition of a set of molecules based on average experimental information. The method was called reconstruction by entropy maximization, since it uses the concept of entropy of information proposed by SHANNON (1948). For that, one should portrait the mixture composition as a probability distribution. Quoting JAYNES (1957) "Information theory provides a constructive criterion for setting up probability distributions based on partial knowledge, and leads to a type of statistical inference which is called the maximum-entropy estimate. It is the least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information." PRESSÉ *et al.* (2013) considers maximum entropy estimates as the only consistent method regarding probability distributions inference.

Following this development, VERSTRAETE *et al.* (2004) proposed a two-step procedure for the molecular reconstruction of vacuum gas oils. In the first step, the stochastic reconstruction framework proposed in TRAUTH *et al.* (1994) is used. Then, with the constructed set of molecules, the reconstruction by entropy maximization was used to calculate mixtures composition.

The development of molecular reconstruction algorithms focuses on the definition of the building blocks of the molecules and the types of probability density functions representing these chemical attributes. Since an optimization step is used, is paramount that the proposed model and its parameters are intimately related to the experimental information available. Then, it is clear that different models should be developed or adapted according to the studied fraction and overall measured properties.

HUDEBINE and VERSTRAETE (2004) proposed a stochastic reconstruction model for light cycle oils, a product of the fluid catalytic cracking unit. The stochastic reconstruction step is used to build a reference mixture. This set of molecules is then used in the reconstruction by entropy maximization algorithm. The authors state that for similar petroleum fractions, once the reference mixture is obtained, only the second step is needed to rebuild different streams. VAN GEEM *et al.* (2007) proposed a method to calculate compositions based on Shannon's information criteria (SHANNON, 1948).

VERSTRAETE *et al.* (2010) extended the two-step procedure proposed in VERSTRAETE *et al.* (2004) to the reconstruction of vacuum residue fractions. The proposed building diagram consisted of 19 distributions with a total of 29 parameters. A genetic algorithm was used as the optimization method. CHARON-REVELLIN *et al.* (2011) used the stochastic reconstruction approach to build a kinetic model for vacuum gas oil hydrotreatment. HUDEBINE and VERSTRAETE (2011) applied

the entropy maximization approach to the reconstruction of fluid catalytic cracking gasolines. DE OLIVEIRA *et al.* (2012) describes a complete modeling methodology based on Monte Carlo sampling for the simulation of the hydrotreating process. DE OLIVEIRA *et al.* (2013) used the stochastic reconstruction algorithm to represent vacuum residues from different origins. The authors show that, if sufficient molecules are built, the stochastic reconstruction step needs to be done only once. Different streams can be represented using the molecular library and the maximum entropy step. DENIZ *et al.* (2017a) used an artificial neural network to reduce the computational demand of the optimization step on stochastic reconstruction models. DENIZ *et al.* (2017b) proposed a building diagram for asphaltenes. They based their choices on the results of compositional spaces reported by McKenna *et. al.* (MCKENNA *et al.*; MCKENNA *et al.*; MCKENNA *et al.*; MCKENNA *et al.*; PODGORSKI *et al.*, 2010a; 2010b; 2013a; 2013b; 2013).

Molecular reconstruction methods are a powerful tool when it comes to build molecular-based models in complex chemical systems, such as the petroleum refining process. In this work, we combined the robustness and flexibility of the stochastic reconstruction algorithm (TRAUTH *et al.*, 1994) with the convenient framework of the structure-oriented lumping method (QUANN and JAFFE, 1992)

3.3 Parameter Inference

As discussed in Section 3.2, one of the most important parts of the stochastic reconstruction algorithms is the estimation of the probability distributions parameters. The literature has been using an optimization approach, based on an objective function. However, the statistical implications of this procedure are neglected. Uncertainty of parameters and predictions or objective function statistical interpretation are not addressed. Moreover, due to the stochastic nature of the model, a procedure that looks for an optimal set of parameters seems counter-intuitive.

Statistical inference can be defined as the task of making conclusions about populations from data. We connect data to the populations using probabilistic models, which in turn are represented by parameters. One popular approach of inference is based on the likelihood principle. The likelihood principle states that all information about the unknown parameters contained in data is represented in the likelihood function (CASELLA and BERGER, 2002). A likelihood is a probabilistic model with data fixed as a function of the unknown parameters. Likelihood ratios measure relative evidences from one set of parameters to another (CASELLA and BERGER; GELMAN *et al.*, 2002; 2014).

Maximum likelihood estimators are a popular method of parameter estimation. In some sense, the literature regarding stochastic reconstruction algorithms uses

this method. Another form of estimation is to calculate the posterior distribution of parameters. This method is known as the Bayesian approach to statistical inference (CASELLA and BERGER; GELMAN *et al.*, 2002; 2014). Bayesian methods transforms a prior distribution into the posterior in light of the observed data using the likelihood function. For that, Bayes theorem is applied. A major advantage of Bayesian methods, especially for complex models, is its natural way to propagate uncertainty.

Stochastic models are a particular class of problems studied in the statistical literature. We shall use the definition proposed in DUTTA *et al.* (2016), and refer to stochastic models as simulator-based models. Simulator-based models are functions that map the model parameters and some random variables to data (DUTTA *et al.*, 2016). Due to the presence of the random variables V, the outputs of the simulator fluctuate randomly even when using the same values of the model parameters (DUTTA *et al.*, 2016). This implies that the likelihood function is intractable, which is a major drawback for maximum likelihood methods.

A specific technique was developed to deal with intractable likelihood problems, namely Approximate Bayesian Computation (ABC). Different algorithms were proposed to solve this problem (BEAUMONT; BEAUMONT *et al.*; BEAUMONT *et al.*; BLUM and FRANÇOIS; BLUM *et al.*; CSILLÉRY *et al.*; DEL MORAL *et al.*; FEARNHEAD and PRANGLE; HICKERSON *et al.*; ROBERT *et al.*; TONI *et al.*; WEGMANN *et al.*; WILKINSON, 2010; 2002; 2009; 2010; 2013; 2010; 2012; 2012; 2006; 2011; 2009; 2009; 2013). For the purpose of this work, we are the first ones to analyze the molecular stochastic reconstruction methods from a Bayesian perspective, estimating the uncertainty of both parameters and predictions.

3.4 Thesis contribution

In this work, we propose a molecular reconstruction algorithm based on both stochastic reconstruction methods (HUDEBINE *et al.*; TRAUTH *et al.*, 2002; 1994) and the structure-oriented lumping method for molecular representation and manipulation (JAFFE *et al.*; QUANN and JAFFE, 2005; 1992). Besides that, we used the reconstruction by entropy maximization approach to calculate mixture composition (HUDEBINE *et al.*, 2002). Our contribution to the literature can be divided into three major topics.

1. An algorithm that combines the robustness and flexibility of the stochastic reconstruction methods with the convenient framework of the structure-oriented lumping molecular representation. We proposed an extension of the chemical attributes of the original SOL method. This new vector of attributes brings

to the structure-oriented lumping a molecular diversity compatible with the stochastic reconstruction algorithms. Besides that, a matrix relating structure-oriented lumping attributes to functional groups is designed.

2. We analyzed the stochastic reconstruction method from a Bayesian perspective. One could argue that approximate Bayesian computation is the most statistical consistent method when it comes to intractable likelihood problems.

3. Stochastic reconstruction algorithms rely on Monte Carlo sample techniques. For that, a large number of samples (molecules) is required to achieve a good representation. Most of the literature on heavy petroleum fractions samples 5000+ molecules. That number may be impractical in some applications. In that sense, we proposed an additional step on the molecular reconstruction algorithms. A non-hierarchical clustering technique to select the best candidates from the ensemble of sampled molecules is proposed. Our clustering method is based on the constraints framework used in the reconstruction by entropy maximization algorithm.

Chapter 4

Molecular Reconstruction Algorithm

In this chapter, we describe the molecular reconstruction algorithm developed in this thesis. Our method is based on the stochastic reconstruction approach proposed by TRAUTH *et al.* (1994) and HUDEBINE *et al.* (2002). Regarding the molecular representation and properties calculation, we developed an extension of the structure-oriented lumping vector proposed by QUANN and JAFFE (1992).

Our work is focused on the heavier fractions of petroleum, especially the vacuum residue, which is the bottom product of the vacuum distillation unit. This fraction is usually sent to a delayed coking unit, or a hydrocracking unit, or a deasphalting unit or is sold as fuel oil (COKER; GRAY; VASSILIOU, 2018; 1994; 2018).

In the development of a molecular reconstruction algorithm, one must follow a logical chain of thought. This chapter is divided in a way that mimics the steps of the design of such algorithms. First, the available experimental data on the considered fraction is defined. This definition serves as an input to the choice of the chemical attributes to be modeled by probability density distributions and in turn the parameters to be estimated. It is paramount that the parameters can be corroborated, at least conceptually, by the experimental data.

Second, the functional forms of the probability density functions and the sampling methodology are defined. Molecular representation, connectivity rules, and properties calculation comes next. In Chapter 5, we discuss the coupling of the stochastic reconstruction with a parameter estimation procedure, the molecular selection by clustering analysis and the composition calculation by entropy maximization.

4.1 Experimental data

The data used to build and test the model came both from the literature and from the database of vacuum residue characterization done at PETROBRAS's research and development center (CENPES).

Specific gravity and average molecular mass. The specif gravity is one of the main properties used to classify oils. It can be seen as an indirect indicator of oil aromaticity, since aromatics have a higher density when compared with saturated molecules of the same molecular mass. In that sense, the relative amounts of molecular types, such as paraffins, naphthenes and aromatics have a great influence on the mixture specific gravity. Average molecular mass is a controversial property for heavy petroleum fractions, mainly because of the aggregation phenomenon taking place in the heavier portion (BODUSZYNSKI; MCKENNA *et al.*; MCKENNA *et al.*, 1987; 2010a; 2010b). However, some of the literature uses this property as input for molecular reconstruction algorithms (DE OLIVEIRA *et al.*; VERSTRAETE *et al.*, 2013; 2010). We used this property to guarantee the reproducibility of literature data.

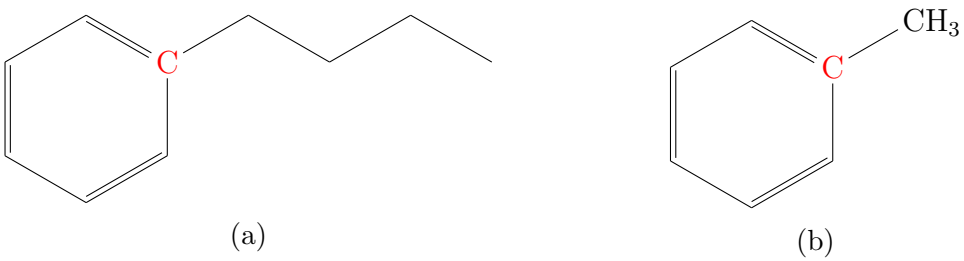
Elemental analysis. The elemental analysis measures the mass percentage of the main atoms present in a petroleum fraction. Carbon, hydrogen, sulfur, nitrogen, and oxygen are the most common results. Similar to the specific gravity, carbon, and hydrogen content is an indirect measure of the oil aromaticity. Besides that, according to the continuity model proposed by ALTGELT and BODUSZYNSKI (1992) and confirmed by MCKENNA *et al.* (2013b), the compositional molecular space extends in terms of aromaticity or carbon-hydrogen ratio. For that, carbon and hydrogen content gives valuable information about molecular types and general molecular structures. Regarding the heteroatoms, the elemental analysis gives only total quantities of this species, giving no information about its functional forms. One should rely on prior chemical knowledge to specify that.

SARA fractions. SARA is an acronym for Saturate, Aromatics, Resins, and Asphaltenes. The method divides the oil into four fractions of the same name. It is based on solubility, so it is a measure of the components polarizability and polarity (FAN *et al.*, 2002). For that, SARA analysis gives information about molecular types, chemical structure and even a general view on functional groups. SARA is also associated with the molecular mass distribution.

^{13}C nuclear magnetic resonance spectroscopy. This method detects a variety of carbon types in the oil fraction analyzed. It is clear how valuable that information is in terms of molecular structure and functional groups. The most common results available for heavy petroleum fractions reports only saturated and unsaturated carbons content. However, in some cases, we have a more detailed

information. The main carbon types detected by NMR are as follows (HASAN *et al.*, 1983):

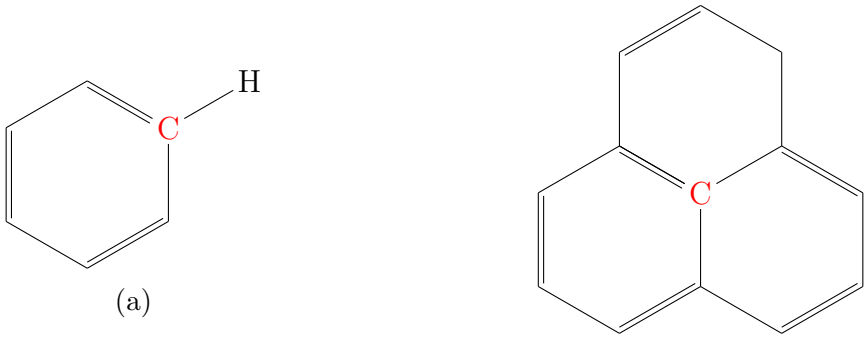
- **Insaturated carbons**
  - Aromatic carbons substituted by an alkyl chain, except if the substituent is a methyl radical.



(a) (b)

Figure 4.1: Carbon types detected by NMR method. (a) Detectable as alkyl substituted aromatic carbon. (b) Detectable as insaturated carbon.

  - Protonated aromatic carbons and internal condensed aromatic carbons.



(a) (b)

Figure 4.2: Carbon types detected by NMR method. (a) Detectable as protonated aromatic carbon. (b) Also detectable as protonated aromatic carbon.

– Peripheral condensed aromatic carbons.

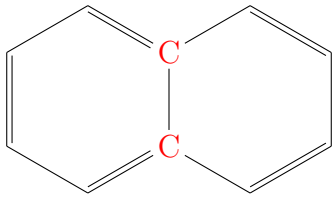


Figure 4.3: Carbon types detected by NMR method. Detectable as insaturated carbons.

- **Saturated carbons**

  – Alpha and beta carbons in a paraffinic chain.

Figure 4.4: Carbon types detected by NMR method. (a) Detectable as α saturated carbon. (b) Detectable as β saturated carbon.

  – Gamma or higher carbons in a paraffinic chain, and naphthenic carbons.

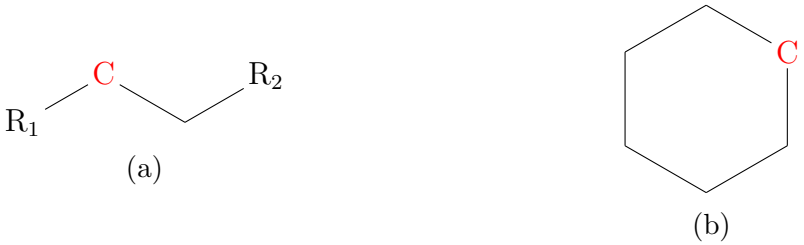


Figure 4.5: Carbon types detected by NMR method. (a) Detectable as γ saturated carbon. (b) Detectable as saturated carbon.

– Branched methyl carbon on a paraffinic chain and terminal methyl carbon on a paraffinic chain.

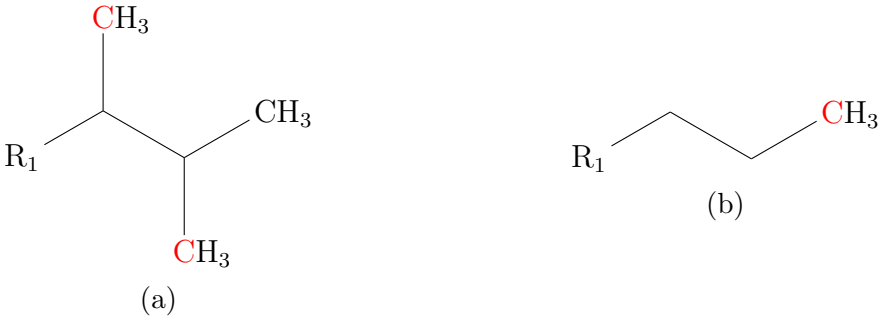


Figure 4.6: Carbon types detected by NMR method. (a) Detectable as branched methyl carbon. (b) Detectable as terminal methyl carbon.

Another useful measure obtained from the carbon nuclear magnetic resonance is the molar percentage of linear alkanes. It is defined as the ratio between CH_2 carbons and the total quantity of carbon atoms.

^{1}H nuclear magnetic resonance spectroscopy. Similar to the carbon type analysis, this method gives information about different types of hydrogen atoms. One can go even further in detailing the molecular structures. The main types of hydrogens detected by this method are as follows (HASAN *et al.*, 1983):

- **Insaturated hydrogens**
  - Aromatic hydrogens.

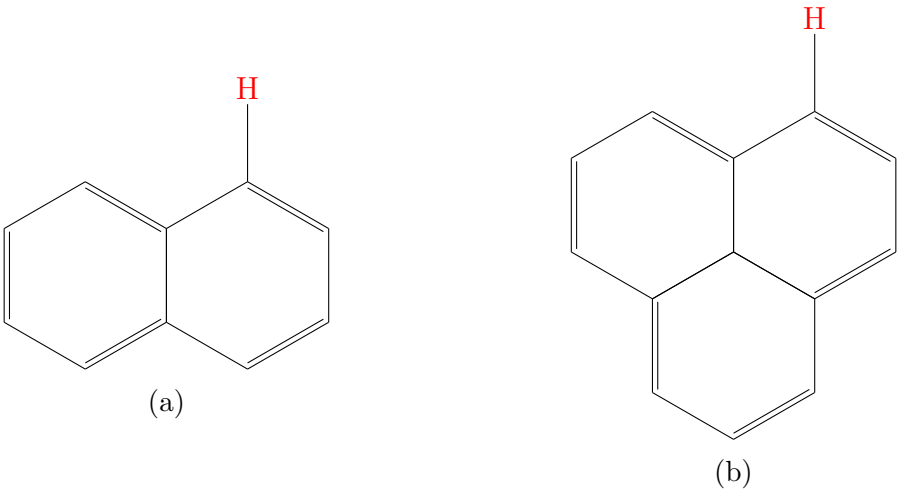


Figure 4.7: Hydrogen types detected by NMR method. (a) Detectable as aromatic hydrogen. (b) Also detectable as aromatic hydrogen.

– Olephinic hydrogens.

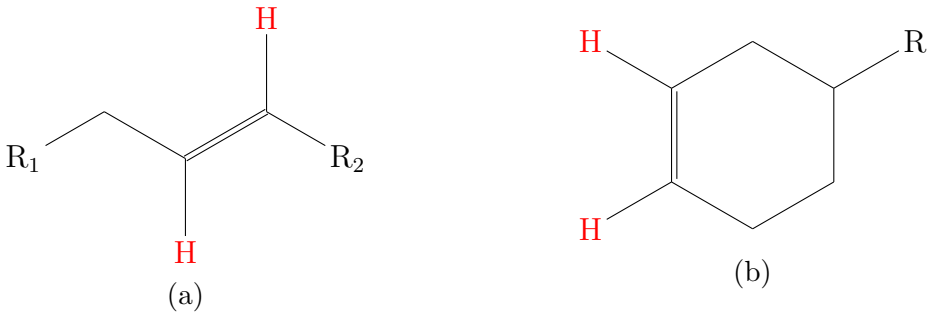


Figure 4.8: Hydrogen types detected by NMR method. (a) Detectable as olephinic hydrogen. (b) Also detectable as olephinic hydrogen.

- **Saturated hydrogens**

  – Hydrogen connected to a carbon in the alpha position of an alkyl substitution in an aromatic ring.

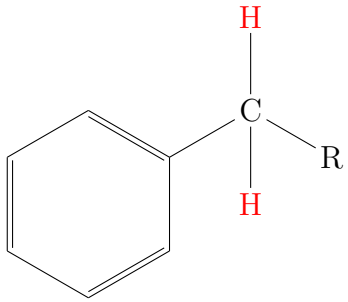


Figure 4.9: Hydrogen types detected by NMR method. (a) Detectable as α hydrogen atoms.

  – Hydrogens connected to terminal or isolated methyl carbons.

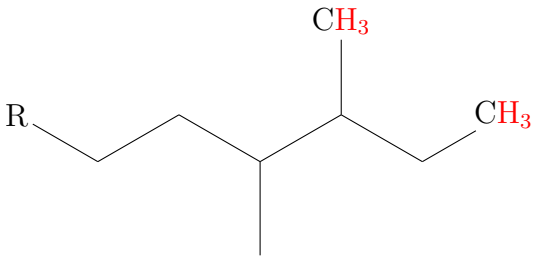


Figure 4.10: Hydrogen types detected by NMR method. Detectable as γ hydrogen atoms. The remaining saturated hydrogen atoms will be detected as β hydrogens.

Although very useful, the detailed nuclear magnetic resonance is not easily available, especially on the refineries day by day operations. For that reason, in this work, we only used the general carbon types results from the NMR, saturated and unsaturated carbon, in the parameter estimation step. However, when available, that information was used to validate the model.

Simulated distillation. This method separates the molecules in terms of their boiling point. The results are reported as the sample vaporized mass percentage for a given temperature. As described by BODUSZYNSKI (1987), the boiling point distribution is highly correlated with the carbon number distribution. In that sense, this information can be used as an estimation of the paraffinc chain length and the number of rings in aromatic cores.

In this section, we described the typical experimental data available for vacuum residues and the relationship between this data and the molecular chemical structure. In Table 4.1 we show a summary of the topics discussed.

Table 4.1: Summary of typical experimental data available and its relationship to chemical structure

Experimental data	**Indirect chemical strucuture information**
Specific gravity	Aromaticity / Relative amounts of molecular types
Elemental analysis	Aromaticity / Functional groups / Heteroatoms abundance
SARA fractions	Relative amounts of molecular types / Polarity / Functional groups / Molecular mass distribution
Carbon and Hydrogen NMR	Functional groups / Molecular structure
Simulated distillation	Carbon number distribution / Paraffinic chain length / Number of aromatic rings in an aromatic core

4.2 Chemical attributes

The basis of the stochastic reconstruction algorithm is the assumption that molecular attributes can be modeled by probability density functions (HUDEBINE *et al.*; KLEIN *et al.*; TRAUTH *et al.*, 2002; 2005; 1994). By chemical attributes, we mean a total number of rings, length of a paraffinic chain and so forth. After combining the typical experimental data presented in Section 4.1 and our prior chemical knowledge, we are able to propose the molecular attributes to be modeled by probability density functions. In this section, we will discuss the models qualitatively. Details about the probability density functions, parameters, and sampling protocols are addressed in Section 4.3.

One common approach when it comes to molecular representation of petroleum fractions is to separate the molecules by molecular types. This has been done both in the analytical chemistry literature (BODUSZYNSKI; LUMPKIN; MEAD; TRESTIANU *et al.*, 1987; 1956; 1968; 1985) and in the molecular reconstruction literature (DE OLIVEIRA *et al.*; DE OLIVEIRA *et al.*; TRAUTH *et al.*, 2013; 2012; 1994). In that sense, the first chemical attribute to be modeled is the molecular type. We used the same molecular types proposed by DE OLIVEIRA *et al.* (2013), which divides the vacuum residue into paraffinics, naphthenics, aromatics, and multicore aromatics. The probability distribution in question has four possible outcomes, matching the molecular types, and its shape defines the relative amounts of such groups. Figure 4.11 is an illustration of the possible outcomes of the distribution. The main advantage of this approach is that once the molecular type is decided, one could treat each group individually according to its main characteristics.

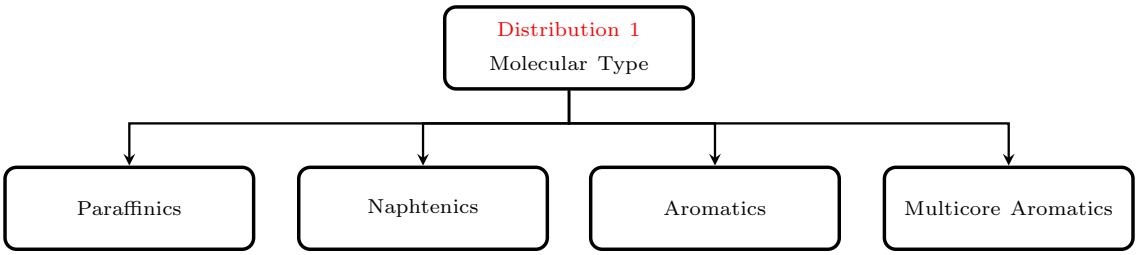


Figure 4.11: First distribution: molecular type.

4.2.1 Paraffinics

A paraffinic molecule contains only aliphatic carbons. Based on experimental evidence, we decided to limit the heteroatoms occurrence to the aromatics and multicore aromatics molecules (GRAY and MCCAFFREY; QIAN *et al.*; QIAN *et al.*; ROSE and FRANCISCO; WALDO *et al.*, 2002; 2001a; 2001b; 1987; 1991). The paraffinic molecule is then defined by the total number of carbons and the level

of branching. These two chemical attributes were modeled by probability density functions. Once is decided to build a paraffinc molecule, one should sample two sequential distributions, the total number of carbons and the level of branching, respectively. Differently from the first distribution, in this case, there is not a finite set of possible outcomes. In Figure 4.12 we show the building diagram of the paraffinic group and an example molecule.

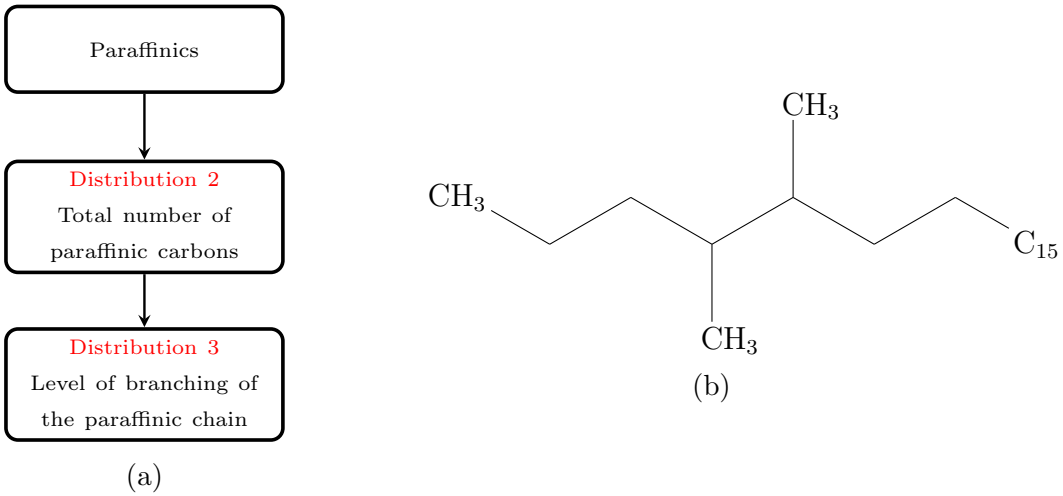


Figure 4.12: Praffinic Molecules. (a) Building diagram. (b) Example molecule - 24 carbons with 2 branches.

4.2.2 Naphthenics

A naphthenic molecule contains at least one cycloparaffin in its structure. It is defined by the following chemical attributes: total number of rings, ring configuration, ring aliphatic substitution, side chain length, and side chain branching level. Similar to the paraffinic molecules, we are not considering the occurrence of heteroatoms in this type of molecule. Once the outcome of the first distribution is a naphthenic molecule, one should sample 5 additional distributions to completely build the molecule. One important thing to notice is that the ring configuration distribution heavily depends on the outcome of the preceding distribution, the total number of rings. This dependence, or conditional probability, is due to the fact that one should consider the available connections to sample the distribution. Also, we model the paraffinic chain length and the side chain length with different distributions. We use the same distribution for the branching level for all types of molecules. In Figure 4.13, we show the proposed building diagram for naphthenic molecules and illustrate a hypothetical naphthenic molecule built with this diagram.

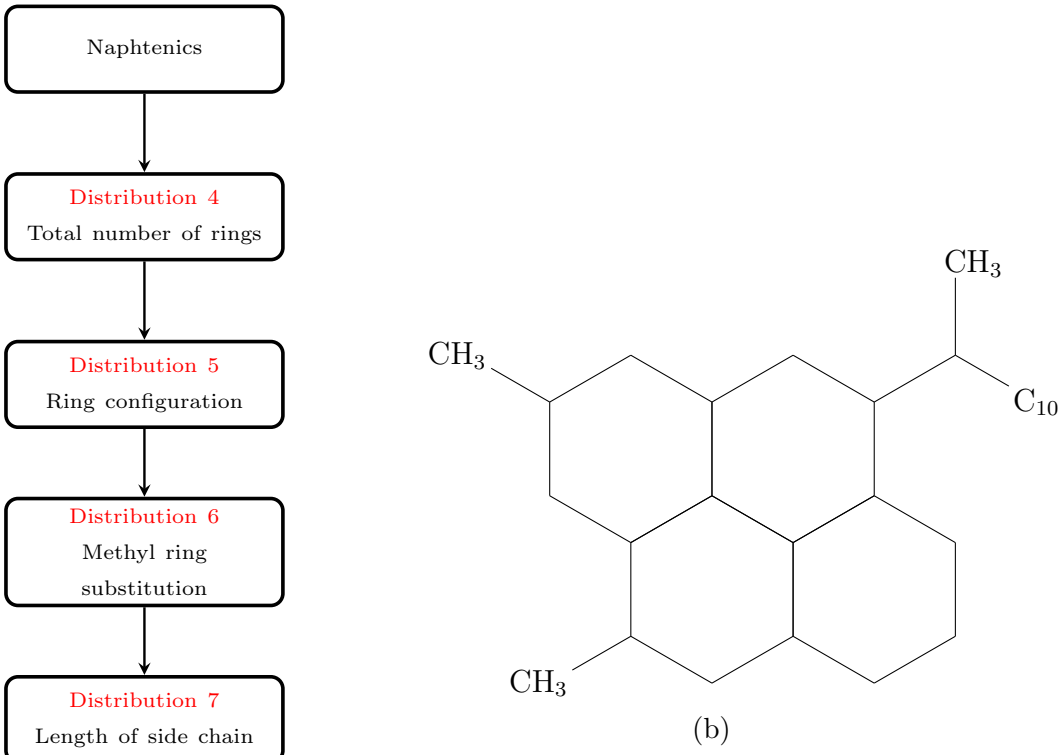


Figure 4.13: naphthenic Molecules. (a) Building diagram. (b) Example molecule - 4 rings, ring configuration *a*, 3 aliphatic ring substitution, 14 carbons in the side chain and 1 branche.

4.2.3 Aromatics

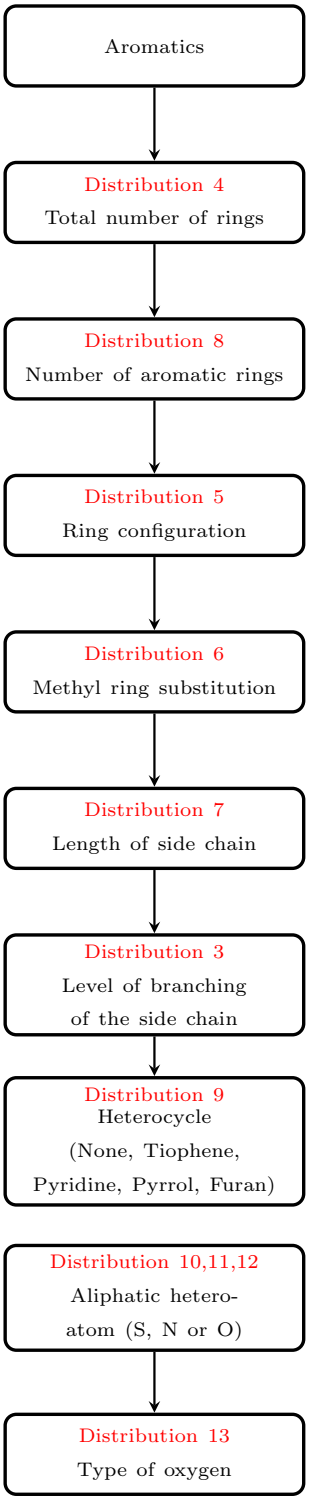
An aromatic molecule contains at least one benzene in its structure. The chemical attributes of the hydrocarbon portion of an aromatic molecule is very similar to that of a naphthenic molecule. For a monocore aromatic, one additional distribution to model the number of benzene rings is included.

As mentioned, the heteroatoms occurrence is restricted to the aromatic molecules. We shall define chemical attributes regarding those species. We divided the heteroatoms into two classes: cyclic and aliphatic. The experimental data available reports only total quantities of these atoms, so we need to rely on prior chemical knowledge to define its functional forms (GRAY and MCCAFFREY; QIAN *et al.*; QIAN *et al.*; ROSE and FRANCISCO; WALDO *et al.*, 2002; 2001a; 2001b; 1987; 1991).

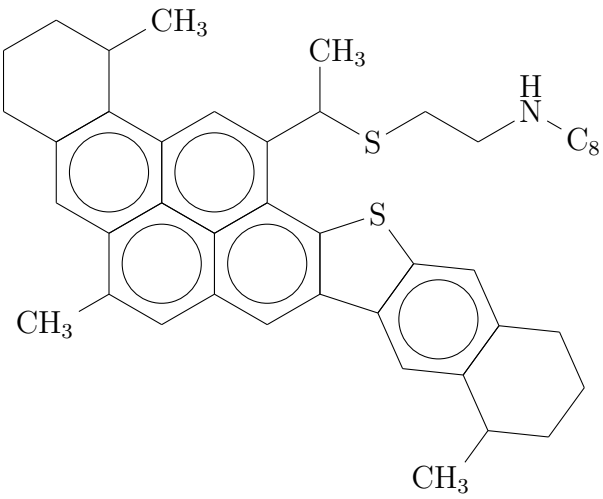
For the cyclic heteroatoms, we proposed a distribution to model the relative amounts of four main groups: thiophene sulfur, pyrrol nitrogen, pyridine nitrogen and furan oxygen. The considered distribution has five possible outcomes: no heterocycle, 1 thiophene, 1 pyrrol, 1 pyridine or 1 furan. For the aliphatic heteroatoms, we used the chemical attributes proposed by DE OLIVEIRA *et al.* (2013), which consists of a distribution to model the probability of a sulfur substitution in an aliphatic chain, a distribution to model the probability of occurrence of a second heteroatom in the aliphatic chain, a distribution to choose between nitrogen and oxygen for the second heteroatom and a distribution to choose the oxygenate function. The main difference with DE OLIVEIRA *et al.* (2013) is in the oxygenate function. In DE OLIVEIRA *et al.* (2013) the choice is between ether and carbonyl functions, in our work, it is between alcohol and aldehyde/ketone functions. In Figure 4.14, we show an example of an aromatic molecule and the aromatics building diagram.

4.2.4 Multicore aromatics

A multicore aromatic molecule is just two or more aromatic cores connected by an aliphatic chain. All the chemical attributes used to model an aromatic molecule are used to build each core of a multicore aromatic molecule. Two additional chemical attributes are necessary. The first one is the number of cores and the second one is the connectivity between cores. Regarding the latter, the distribution decides how many connections a core will make and the type of connection (aromatic-aromatic, aromatic-naphthenic, naphthenic-naphthenic). The length of the aliphatic bridge between two cores uses the same distribution used for side chain length. In Figure 4.15, we show the multicore building diagram and one example multicore aromatic molecule.

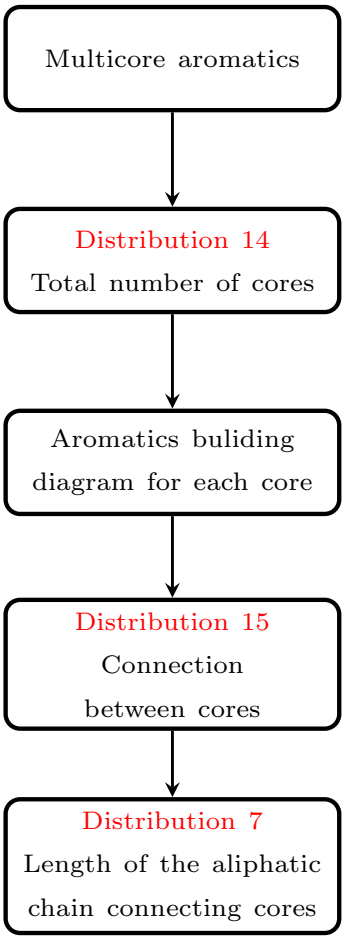


(a)

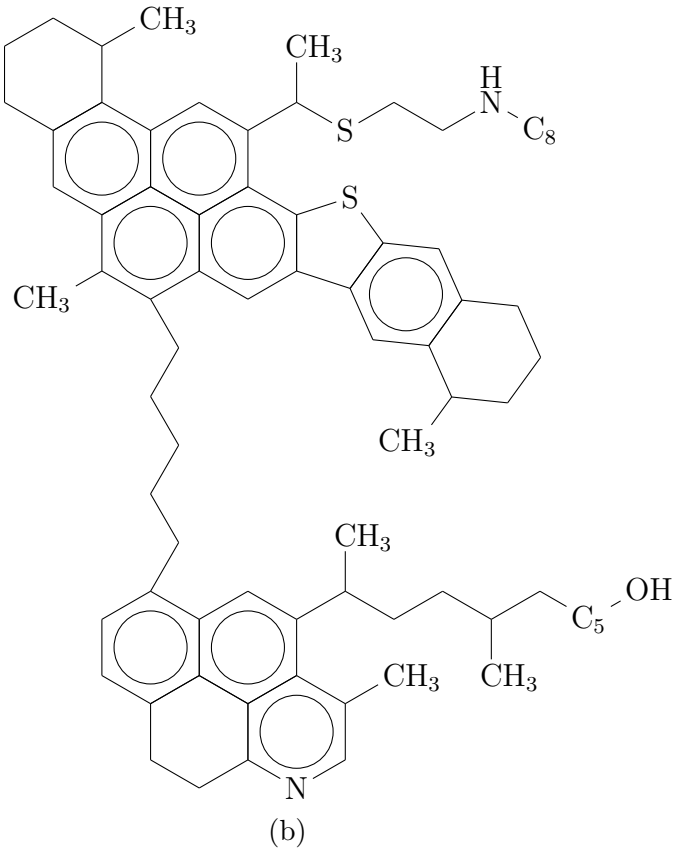


(b)

Figure 4.14: (a) Aromatic building diagram. (b) Example aromatic molecule: 7 total rings, 5 benzene rings, ring configuration *b*, 4 methyl rings substitution, 16 carbons on the side chain, 1 branche, 1 thiophene, 1 aliphatic sulfur, 1 aliphatic nitrogen.



(a)



(b)

Figure 4.15: (a) Multicore aromatics building diagram. (b) Example multicore aromatic molecule: 2 cores.
Core 1: 7 total rings, 5 benzene rings, ring configuration *b*, 4 methyl rings substitution, 16 carbons on the side chain, 1 branche, 1 thiophene, 1 aliphatic sulfur, 1 aliphatic nitrogen.
Core 2: 4 total rings, 3 benzene rings, ring configuration *c*, 2 methyl rings substitution, 13 carbons on the side chain, 2 branches, 1 pyridine, 1 aliphatic oxygen (alcohol).
Core connections: 5 carbons, connection type 1 (aromatic-aromatic).

4.3 Probability density functions, parameters and sampling protocol

In Section 4.2, we focused on the definition of the chemical attributes to be modeled by probability density functions. This definition was based on both experimental data and prior chemical knowledge. In this section, we define the functional forms for the proposed distributions and the sampling protocol.

Probability density functions

When evaluated at some input value, a probability density function returns the probability of a random variable to assume that input value. In our case, the random variables are chemical attributes. Probability density functions can be discrete or continuous in terms of the random variable. Discrete probability density functions are usually referred to as probability mass functions (CASELLA and BERGER, 2002). To be a normalized PDF, a function $p(x)$ must satisfy the following conditions:

$$p(x) > 0, \tag{4.1}$$

$$\int_{-\infty}^{+\infty} p(x)dx = 1 \quad (continuous), \tag{4.2}$$

$$\sum_{i=0}^{N} p(x_i) = 1 \quad (discrete). \tag{4.3}$$

Cumulative density functions

When evaluated at some input value,x_i, a cumulative density function returns the probability of a random variable to be less or equal to that input value. When dealing with continuous random variables, the probability density function can be obtained as the derivative of the cumulative density function. The cumulative distribution function can be represented as follows:

$$P(x_i) = \int_{-\infty}^{x_i} p(x)dx \tag{4.4}$$

Monte Carlo sampling

In order to build molecules in the stochastic reconstruction framework, one should sample from the proposed probability distributions. After that, we can assemble the outcomes in terms of chemical structure, as described in Section 4.2. The least biased way to do that, is to generate random samples from those distributions using

a Monte Carlo sampling protocol. For that, we use the concept of equivalent random sequences, defined as follows:

$$\int_{-\infty}^{x_i} p_1(x)dx = \int_{-\infty}^{y_i} p_2(y)dy. \tag{4.5}$$

In Equation 4.5, one can see that equivalent random sequences are the ones that generates the same cumulative probabilities for different distributions. By generating uniformely distributed random numbers between 0 and 1, we are able to transform that sequence into any distribution considered. A uniform distribution between 0 and 1 has the following propertie:

$$\int_{-\infty}^{x_i} p_1(x)dx = x_i. \tag{4.6}$$

Equation 4.5 becomes:

$$x_i = \int_{-\infty}^{y_i} p_2(y)dy. \tag{4.7}$$

To generate a sequence of random numbers y_i from any distribution $P_2(y)$, we only need to encounter the value y_i that has the cumulative probability in the considered distribution equivalent to the uniformly generated number x_i.

Discretization, truncation and renormalization

Chemical attributes, when modeled by probability density functions, can be seen as discrete random variables. A molecule can not have 10.5 carbons. Therefore, when using common continuous distributions to model chemical attributes one should discretize them first. The discretization can be done by considering ranges of cumulative probabilities instead of absolute values. For instance, imagine a distribution where the probability of a random variable to be less or equal 10 is 0.8 and the probability of the same variable to be less or equal 10.9 is 0.82. In this case, probabilities ranging from 0.8 to 0.82 are associated with the value 10 of the random variable.

Besides being discrete values, chemical attributes are also finite. When using probability density functions that covers all positive real numbers, one should consider using a truncated form. In this work, we used the truncation criteria proposed by TRAUTH *et al.* (1994). This criteria consists of truncating a distribution in the value of the random variable x_{i+1} that contributes to the total cumulative probabilitiy with less then 0.1 % in a relative basis, as follows:

$$\frac{\int_{-\infty}^{x_{i+1}} p(x)dx - \int_{-\infty}^{x_i} p(x)dx}{\int_{-\infty}^{x_i} p(x)dx} \leq 0.1\,\% \tag{4.8}$$

Clearly, after truncation, one should normalize the function to guarantee that the probabilities sum to one.

Conditional probability

One final important topic to be discussed is the conditional probability in the sampling protocol. Some distributions heavily depend on the outcomes of the preceding distributions, to the point that they need to be rebuilt at every sampling step. For example, the number of benzenes can only be as high as the total number of rings. The distribution modeling the number of benzenes has to be truncated in a different point every time the preceding distribution is sampled. Another example is the ring configuration. We use this distribution to choose between different points to connect a ring. In Figure 4.16, we show the decision process in the construction of a 4 ring aromatic core. One can see that every time the distribution is sampled it has to be rebuilt, since the possible outcomes changes.

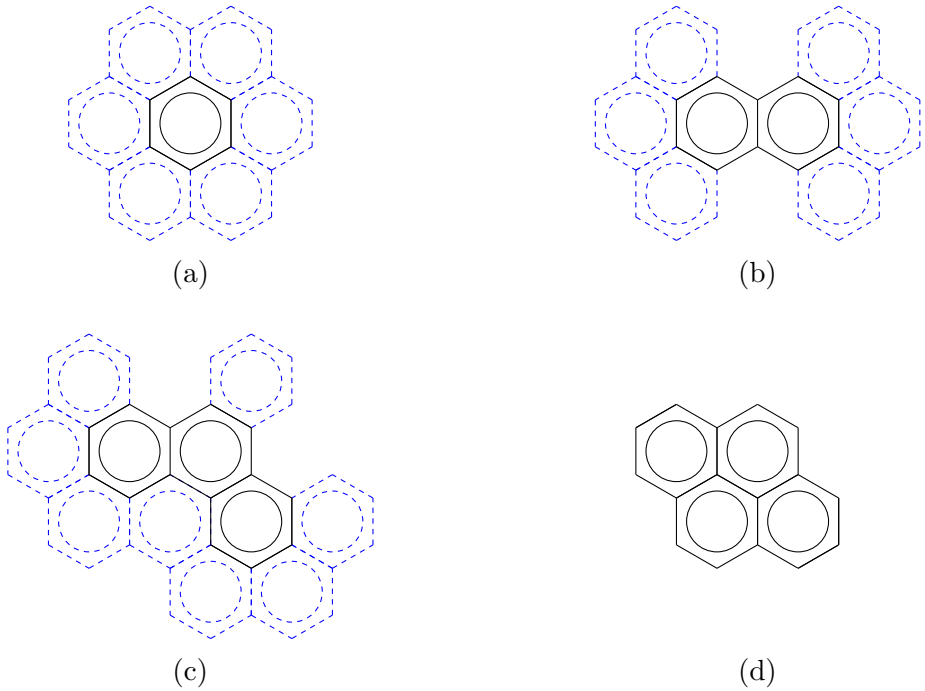


Figure 4.16: Example of ring connection decision process. Dashed blue lines represents possible entrance points for the next ring. The distribution needs to be rebuilt in every step of the core construction. (a) Step 1 : 6 possible outcomes, (b) Step 2 : 6 possible outcomes, (c) Step 3 : 9 possible outcomes (d) Final molecule.

After discussing the general concepts of probability density functions and sampling methodologies, we are able to define the functional forms of the distributions described in Section 4.2. Some of the distributions cannot be estimated by experimental data. In such cases, we define the distributions prior to the observation of experimental data.

Distribution 1 - Molecular type

The first distribution has only four possible outcomes: paraffins, naphthenes, aromatics and multicore aromatics. It is hard to model this kind of distribution with a fixed functional form. One possible approach is to not define any standard shape. For that, one should consider each probability associated with each of the possible outcomes to be an unknown parameter. Considering the available experimental data, properties like specific gravity and SARA fractions are good measures of relative amounts of molecular types, as discussed in Section 4.1. In that sense, we should be able to estimate those parameters. Since probabilities should sum up to one, the number of parameters for this kind of distribution is the number of possible outcomes minus one. We shall use the definition proposed by DE OLIVEIRA *et al.* (2013) that calls this type of distribution by histogram. In Figure 4.17, we show an example of this distribution.

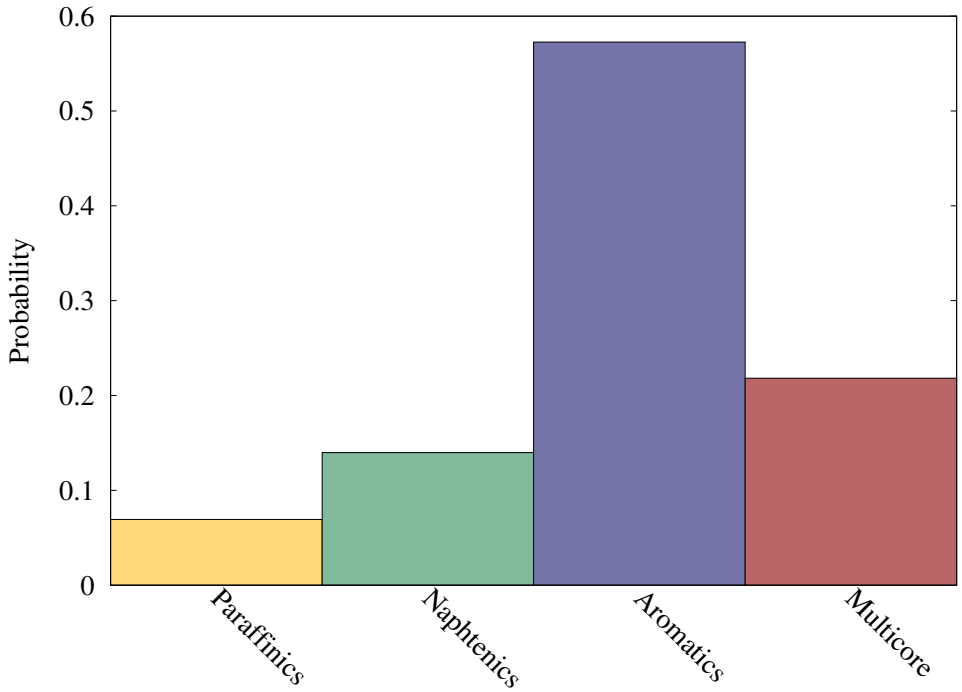


Figure 4.17: Example of a distribution for the molecular type

Distribution 2 - Lenght of a paraffinic chain

Experimental evidence shows that the distribution of the number of carbons has a similar shape when compared to the boiling point distribution. BODUSZYNSKI (1987) proposed a gamma distribution to represent this attribute. Since we are considering the simulated distillation as available experimental data, we should be able to estimate the distribution parameters. The gamma distribution is very flexible in its shape, depending on the values of the parameters it can be close to a normal distribution or an exponential distribution. However, with the intention of reducing the total number of parameters, we used a particular case of the gamma distribution: the chi-squared distribution. It's functional form can be seen in Equation 4.9

$$p(\chi^2, \nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}(\chi^2)^{[\nu/2-1]}e^{(-\chi^2/2)}, \tag{4.9}$$

where χ is the random variable or the chemical attribute in our case, ν is defined as the degrees of freedom and the only parameter of the distribution and Γ is the gamma function. In Figure 4.17, we show an example of such a distribution.

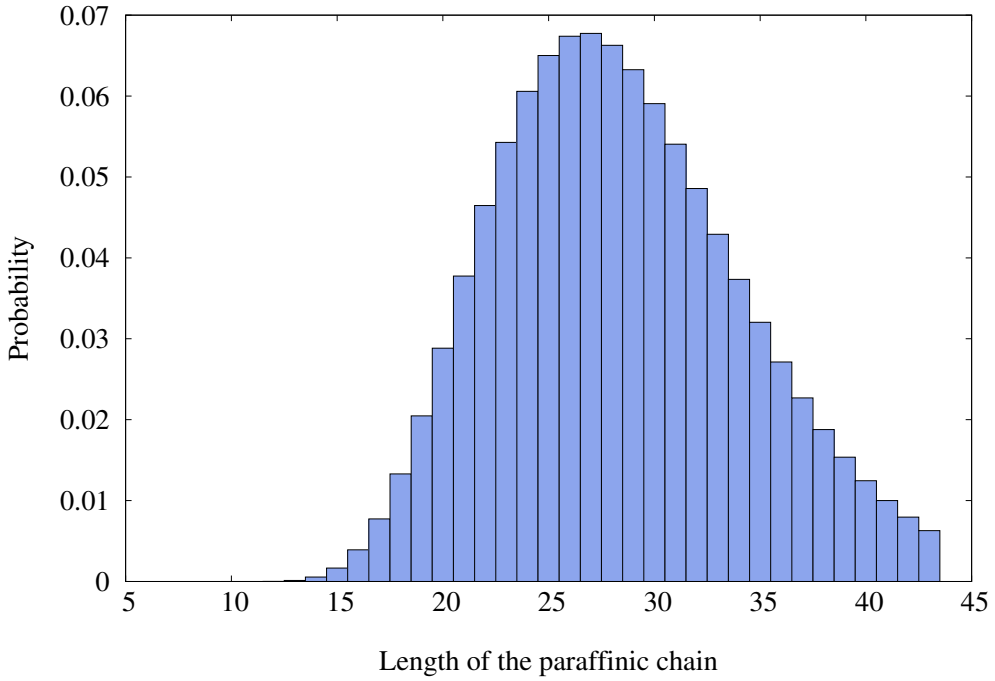


Figure 4.18: Example of a distribution for the length of paraffinic chain

Distribution 3 - Level of branching in an aliphatic chain

The experimental data capable of defining this distribution is the quantity of branched methyl carbons detected by nuclear magnetic resonance. However, as men-

tioned in Section 4.1, we did not consider this data in the parameter estimation step. The distribution was defined by means of pure chemical knowledge (ALTGELT and BODUSZYNSKI; BODUSZYNSKI; BODUSZYNSKI; BODUSZYNSKI and ALTGELT; HIRSCH and ALTGELT; LUMPKIN; MCKENNA *et al.*; MCKENNA *et al.*; MCKENNA *et al.*; MCKENNA *et al.*; PODGORSKI *et al.*, 1992; 1987; 1988; 1992; 1970; 1956; 2010a; 2010b; 2013a; 2013b; 2013). We considered a maximum number of 4 branches per aliphatic chain. Since we do not have much information, we chose a uniform distribution from 0 to 4 to model this attribute. All outcomes of the distribution are equally probable. Some outcomes may have its probability changed conditional to the preceding distribution. For instance, a 3 carbon paraffinic molecule cannot have any branches. In this case, all the probabilities are zero.

Distribution 4 - Total number of rings

The total number of rings is related to the total number of carbons and in turn to the boiling point distribution. We can also relate it to the carbon-hydrogen ratio. Similar to the length of paraffinic chain distribution, we used the chi-squared distributions to model this attribute. In Figure 4.19, we show an example of the total number of rings distribution.

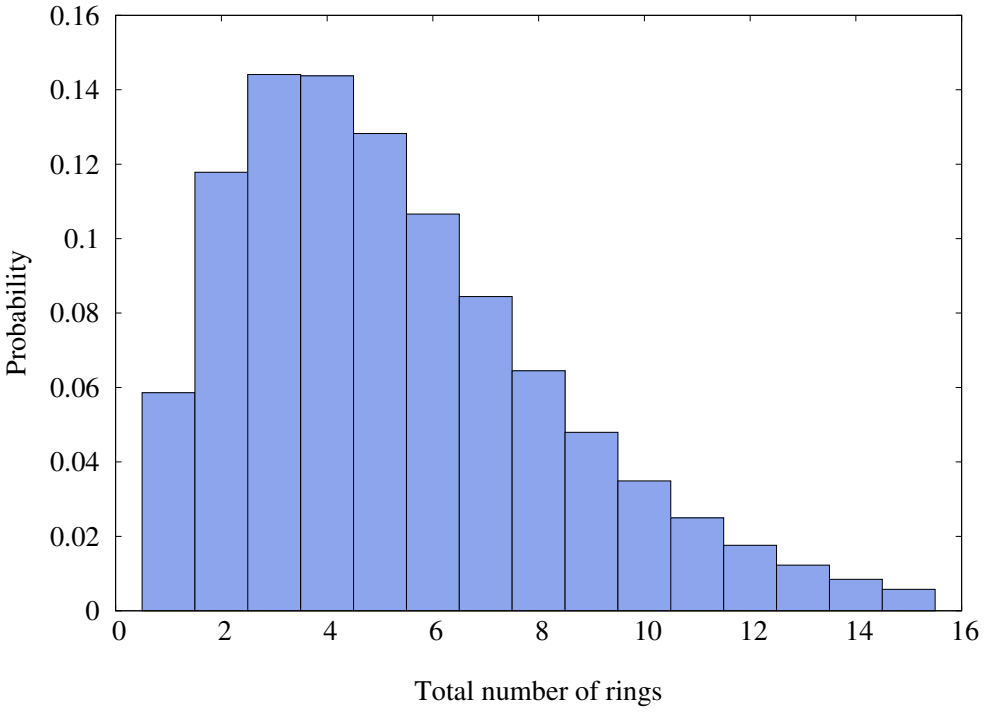


Figure 4.19: Example of a distribution for the total number of rings

Distribution 5 - Ring configuration

Ring configuration is more related to specific carbon types rather than general properties, such as specific gravity and carbon-hydrogen ratios. Specific carbon types can be obtained by the detailed nuclear magnetic resonance spectroscopy described in Section 4.1. However, we did not consider this type of data in the parameter estimation. In that sense, we used a uniform distribution based on every possible outcomes in terms of ring connection. In Figure 4.16, we illustrate the sampling steps to the final ring configuration. In every step, every possible connection (dashed blue lines in Figure 4.16) is equally probable.

Distribution 6 - Methyl ring substitution

Another distribution that requires detailed nuclear magnetic resonance spectroscopy data. For this distribution, we used the same approach used for distribution 3, branching level. An uniform distribution ranging from 0 to 4, 0 meaning no substitution. Once again, the probabilities may change depending on connections site availability.

Distribution 7 - Lenght of the side chain

One can think about the side chain as an increment to the total number of rings in terms of carbon numbers. In that sense, this chemical attribute greatly influences the heavier portion of the boiling point distribution curve. Based on the work of BODUSZYNSKI (1987), the boiling pointing curve extends exponentially towards the heavier portions. To model this attribute, we choose the exponential function, defined as follows:

$$p(t, \beta) = \frac{\exp\left(\frac{-t}{\beta}\right)}{\beta}, \tag{4.10}$$

where t is the random variable associated with the exponential distribution and θ is the parameter defining its shape. In Figure 4.20, we show an example of the exponential representation of the side chain length.

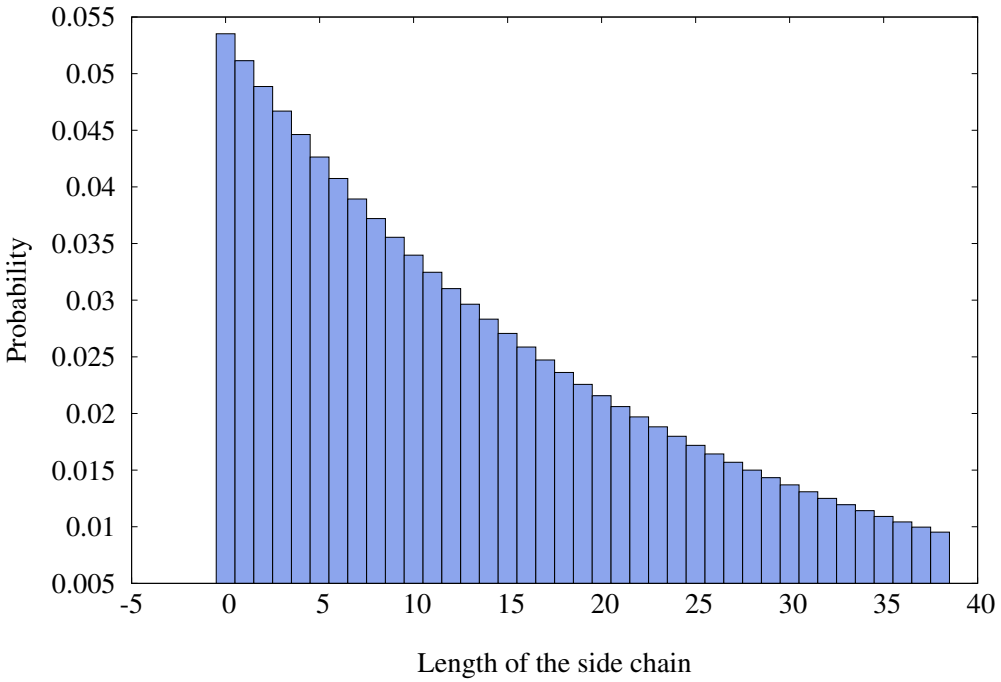


Figure 4.20: Example of a distribution for the length of the side chain

Distribution 8 - Number of benzenes

This chemical attribute greatly influences the carbon to hydrogen ratio, specific gravity and SARA fractions relative amounts. Similar to the side chain length, this is an incremental attribute with a greater influence in the heavier portion of the carbon number distribution. For that, we used the exponential function defined in Equation 4.10 as the distribution functional form. In Figure 4.21, we show an example of distribution for this chemical attribute.

Distribution 9 - Type of heterocycle

Experimental data only report total amounts of heteroatoms. The definition of their functional forms was based on pure chemical knowledge (GRAY and MCCAFFREY; QIAN *et al.*; QIAN *et al.*; ROSE and FRANCISCO; WALDO *et al.*, 2002; 2001a; 2001b; 1987; 1991). This distribution has only five possible outcomes: no heterocycle, 1 thiophene, 1 pyrrol, 1 pyridine or 1 furan. Similar to the molecular type distribution we decided to use a histogram (free-shaped distribution) as the probability density function. One important observation is that, since we measure only total amounts, we can not differentiate between pyrrolic and pyridinic nitrogen. In that sense, we estimate one parameter that controls the amounts of the sum of both functional forms, each one being equally probable. In Figure 4.22, we show an example of this type of distribution.

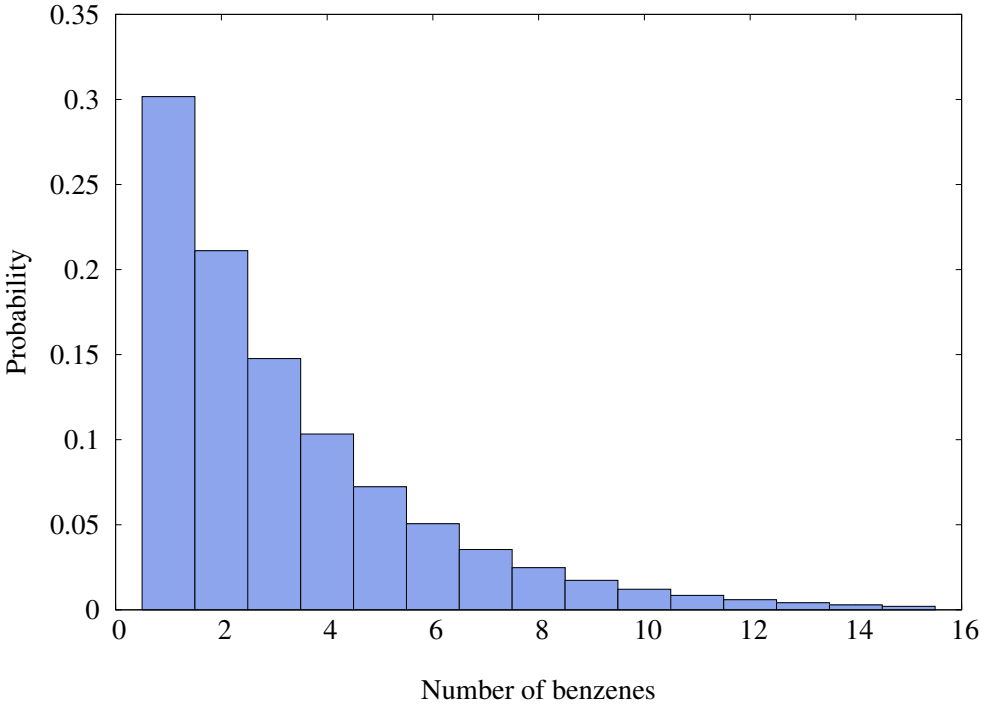


Figure 4.21: Example of a distribution for the number of benzenes

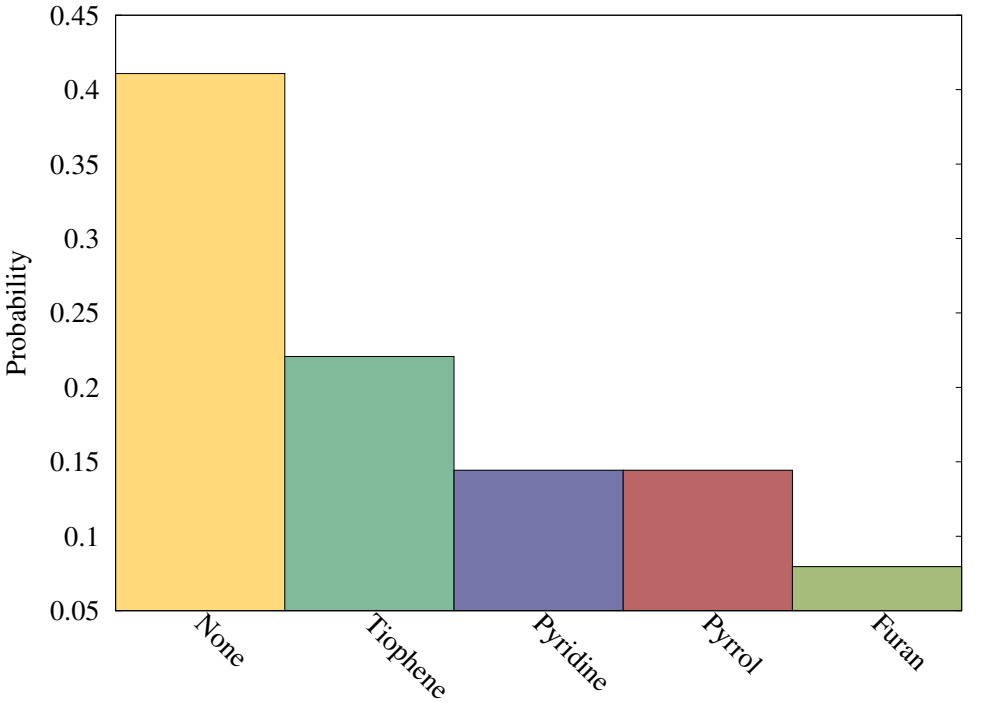


Figure 4.22: Example of a distribution for the type of heterocycle

Distributions 10,11,12 and 13 - Aliphatic heteroatoms and oxygenate function

We used the same framework proposed by DE OLIVEIRA *et al.* (2013) to model these chemical attributes. It consists of three distributions, each one with two possible outcomes. The first one decides for a sulfur, the second one for a second heteroatom and the third one chooses between nitrogen and oxygen. Based on the low number of possible outcomes, we used a histogram type distribution. Since we only have total amounts of heteroatoms as experimental data, chances are that these parameters are correlated to those of distribution 9. Nevertheless, we kept them in the parameter estimation step. Regarding the oxygenate function, we considered both options equally probable.

Distribution 14 - Number of cores

Similar to the length of the side chain, this chemical attribute is related to the heavier portion of the boiling point distribution. As previously discussed, the exponential function defined in Equation 4.10 is suitable to model these types of attributes. In Figure 4.23, we show an example for this distribution.

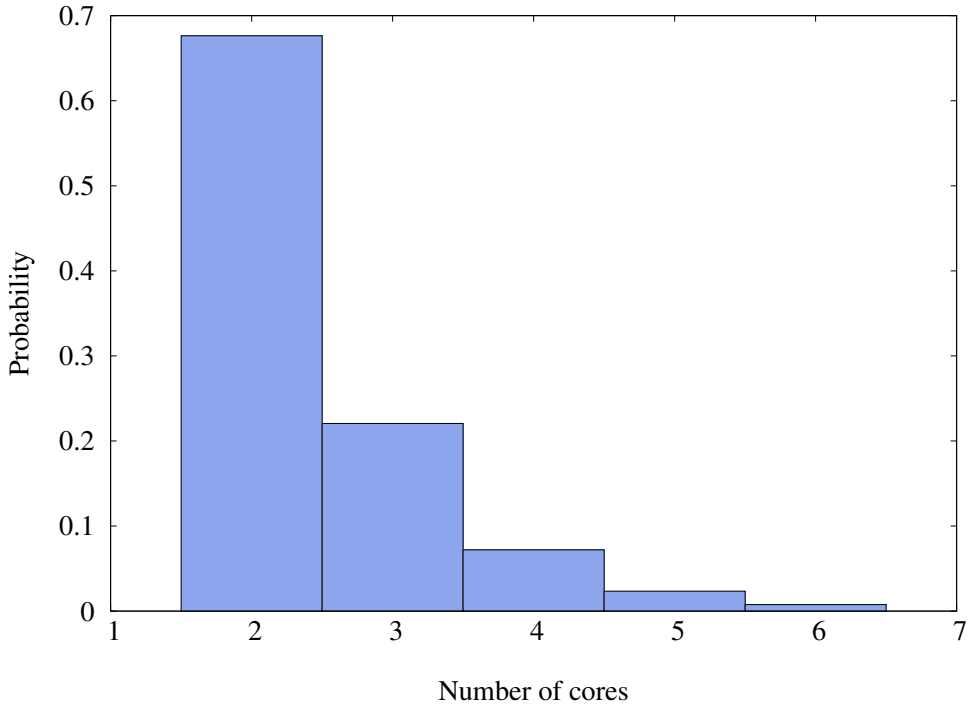


Figure 4.23: Example of a distribution for the number of cores

Distribution 15 - Connection between cores

To fully estimate this distribution, one would need detailed nuclear magnetic resonance. Similar to distributions 3, 5 and 6, we used a uniform distribution to model both the number of connections and type of connection. For the first, we considered a maximum of 4 connections per core, each outcome being equally probable. The same thinking was applied to the connections type distribution.

In this section, we discussed the general concepts regarding modeling chemical attributes with probability density functions. Besides that, we defined the functional forms of the distributions and whether or not the parameters can be estimated by experimental data. For dealing with distributions with no experimental data available, we used the uniform functional form. By doing that, we preserve our prior chemical knowledge in the least biased way. In Table 4.2, we show a summary of the distributions defined in this section.

Table 4.2: Summary of distributions functional forms, chemical attributes and parameters to be estimated

Chemical attribute	**Functional form**	**Available data?**	**Parameters**
Molecular type	Histogram	yes	3
Lenght of a paraffinic chain	Chi-squared	yes	1
Level of branching	Uniform	no	-
Total number of rings	Chi-squared	yes	1
Ring configuration	Uniform	no	-
Methyl ring substitution	Uniform	no	-
Lenght of the side chain	Exponential	yes	1
Number of benzenes	Exponential	yes	1
Type of heterocycle	Histogram	yes	3
Aliphatic sulfur	Histogram	yes	1
Another aliphatic heteroatom	Histogram	yes	1
Aliphatic nitrogen or oxygen	Histogram	yes	1
Aliphatic oxygen function	Uniform	no	-
Number of cores	Exponential	yes	1
Connection between cores	Uniform	no	-

4.4 Molecular representation

In Section 4.3, we described the sampling methodology and the functional forms of the distributions representing the chemical attributes. For each sample of the building diagram, one molecule is built. The stochastic reconstruction algorithm consists of taking N samples from the building diagram. These samples represent the molecules in the hypothetical mixture. At this point, we consider that every molecule has the same mole fraction in the mixture, equal to $1/N$.

In this section, we describe the methodology used to represent and storage the molecules generated by the Monte Carlo sampling procedure.

Structure-oriented lumping - monocore molecules

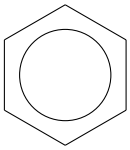
The structure-oriented lumping method, as proposed by QUANN and JAFFE (1992), represents individual hydrocarbons molecules as vectors of 22 structural increments. These structural increments are shown in Figure 4.24.

A6	**A4**	**A2**	**N6**	**N5**	**N4**	**N3**	**N2**	**N1**	**R**	**br**	**me**	**IH**	**AA**	**NS**	**RS**	**AN**	**NN**	**RN**	**NO**	**RO**	**KO**

Figure 4.24: Structural increment attributes of the structure-oriented lumping method (QUANN and JAFFE, 1992)

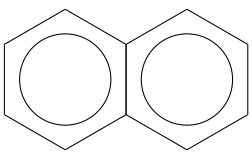
A mixture of molecules is then represented by a matrix. Each row represents one molecule, and the columns are the structural attributes shown in Figure 4.24. The nature of the structural increments are defined as follows (QUANN and JAFFE, 1992):

- **A6:** A six carbon aromatic ring.



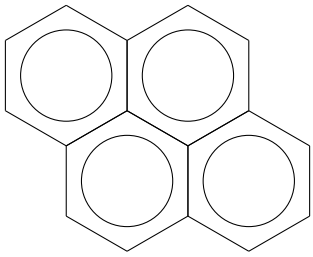
A6	**A4**	**A2**	**N6**	**N5**	**N4**	**N3**	**N2**	**N1**	**R**	**br**	**me**	**IH**	**AA**	**NS**	**RS**	**AN**	**NN**	**RN**	**NO**	**RO**	**KO**
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **A4:** A four carbon aromatic ring increment. It has to be attached to either an A6 or another A4.



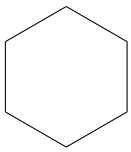
A6	**A4**	**A2**	**N6**	**N5**	**N4**	**N3**	**N2**	**N1**	**R**	**br**	**me**	**IH**	**AA**	**NS**	**RS**	**AN**	**NN**	**RN**	**NO**	**RO**	**KO**
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **A2:** A two carbon aromatic ring increment. Results in a pericondensed multiring structure as in pyrene.

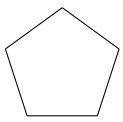


A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **N6 and N5:** Six and five carbon naphthenic rings. Similar to the A6 increment, they can exist independently as cyclohexane and cyclopentane.

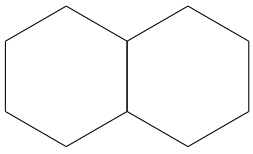


A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

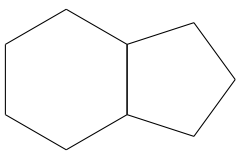


A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

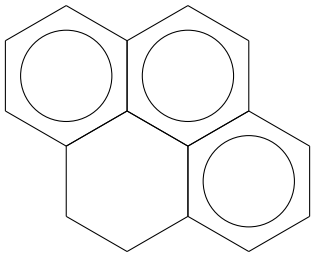
- **N4, N3, N2, and N1:** Four three two and one carbon naphthenic ring that must be attached to other naphthenic or aromatic ring structure.



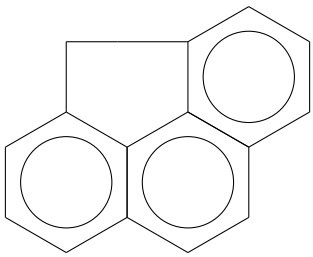
A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0



A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

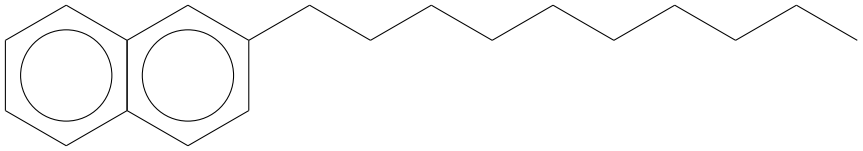


A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0



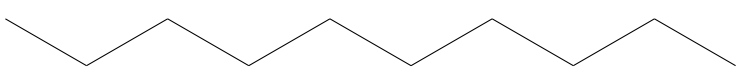
A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

- **R:** Carbon number of the total alkyl substitution in a ring structure or the carbon number of aliphatic molecules. the R increment adds $-CH_2-$ groups.

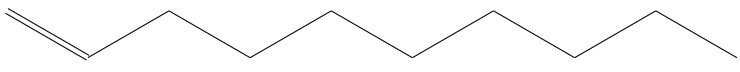


A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	1	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0

- **IH:** Incremental hydrogen to specify the degree of unsaturation of molecules. It adds two hydrogen atoms to the stoichiometry of a molecule. If no rings are present, IH = 1 for paraffins, IH= 0 for monoolefins, and IH = -1 for diolefins. If there are naphthenic rings present, IH=-1 indicates a cycloolefin.

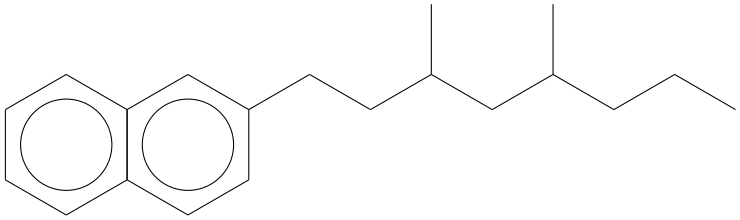


A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
0	0	0	0	0	0	0	0	0	10	0	0	1	0	0	0	0	0	0	0	0	0



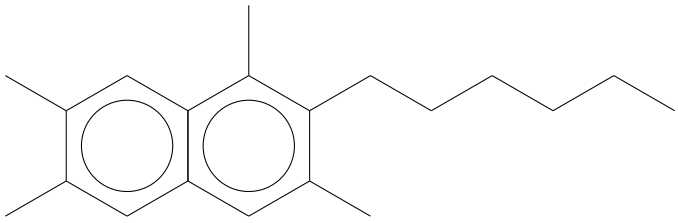
A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0

- **br:** Indicates the number of branch points on the alkyl side chain R or on a paraffin or olefin. The br group contributes no hydrogen or carbon to the stoichiometry of the molecule.



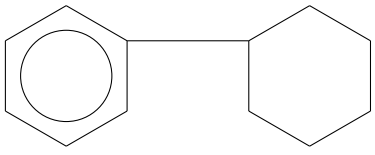
A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	1	0	0	0	0	0	0	0	10	2	0	0	0	0	0	0	0	0	0	0	0

- **me:** Specifies the number of carbons of the total alkyl structure R which are attached as methyl groups to the carbon atoms on aromatic or naphthenic rings of a molecule. The group me also does not contribute carbon or hydrogen to the stoichiometry of the molecule.



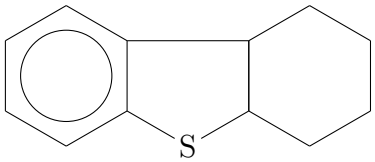
A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	1	0	0	0	0	0	0	0	10	0	4	0	0	0	0	0	0	0	0	0	0

- **AA:** The biphenyl bridge between any two nonincremental rings (A6, N6, or N5). AA contributes no carbon to the structure but eliminates two hydrogen atoms to form the bridge between rings.

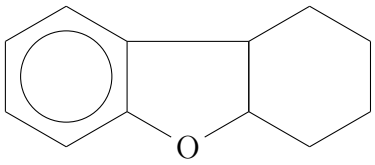


A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

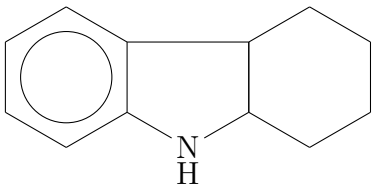
- **NS, NN, and NO:** A sulfur, nitrogen, or oxygen located in a naphthenic ring or paraffin and bound to two carbon atoms. NS, NN, or NO replaces a CH_2 methylene unit with an S atom, an $N-H$ group, or an oxygen atom in the structure, respectively.



A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	0	0	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0

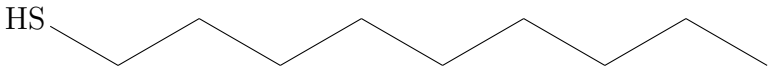


A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0

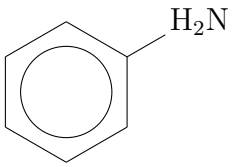


A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0

- **RS, RN, and RO:** A sulfur atom, nitrogen -NH- group, or oxygen atom inserted between a carbon and hydrogen atom to form a mercaptan, amine, or alcohol group, respectively.

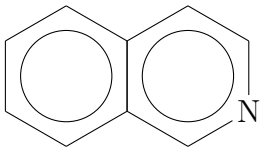


A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
0	0	0	0	0	0	0	0	0	10	0	0	1	0	0	1	0	0	0	0	0	0



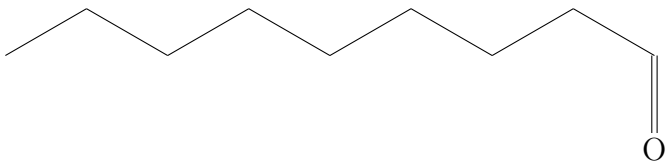
A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

- **AN:** A nitrogen group substitution for carbon in an aromatic ring as in pyridine.



A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

- **KO:** A ketone or aldehyde group where a $-\mathbf{CH_2}-$ is replaced by $> \mathbf{C = 0}$.



A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Structure-oriented lumping - multicore molecules

In JAFFE *et al.* (2005), an extension of the structure-oriented lumping method was proposed. A new vector of structural attributes was proposed to better represent vacuum residues. Moreover, a methodology to represent multicore molecules was presented. In this work, we used a methodology based on JAFFE *et al.* (2005) original proposition to represent multicore molecules.

First of all, the molecular matrix needs another dimension. Now, each row represents a core of a molecule, the columns still represent the structural attributes and the third dimension represents the different cores of the molecule. We added two additional attributes to the structure vector, as shown in Figure 4.25.

A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO	Cn	Nc

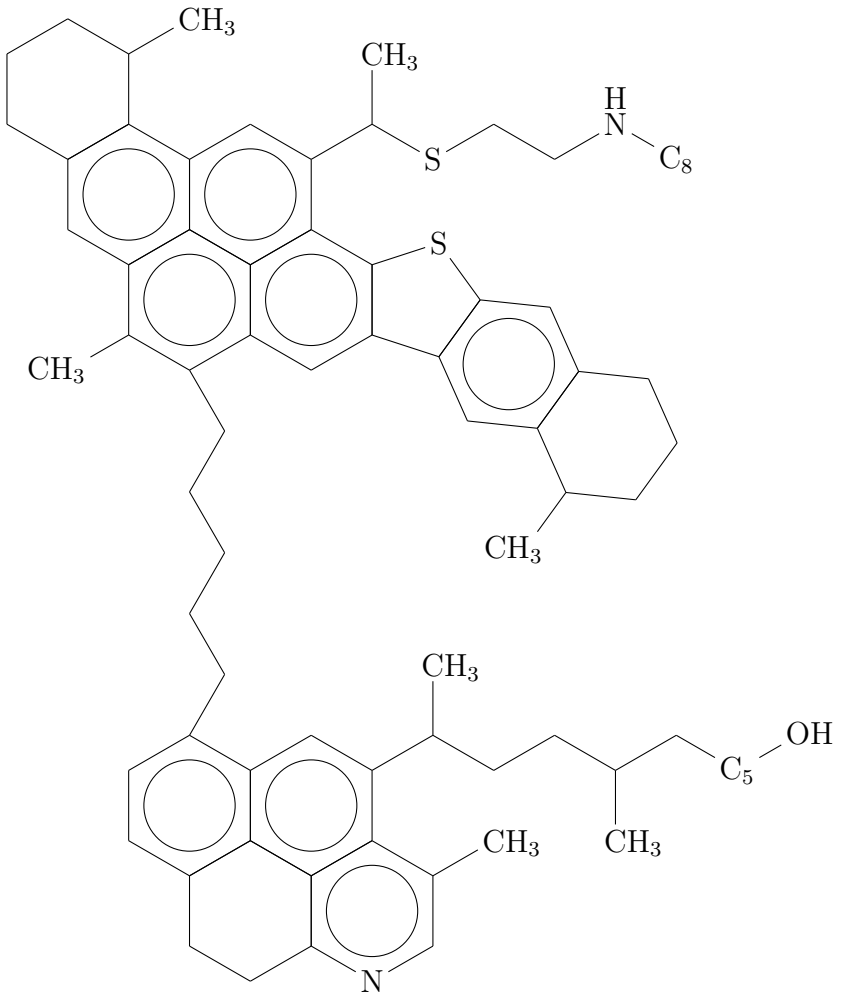
Figure 4.25: Augmented vector of structural attributes to represent multicore molecues. The additional Cn and Nc attributes are used to represent multicore molecules.

The additional attributes are defined as follows:

- **Cn:** Connectivity between cores. As described in Section 4.3, we considered a maximum of four connection per core and three types of connections. This attribute is a code consisting of eight digits. The first four digits defines to which cores the current core connects. The last four digits define the type of connection.

- **Nc:** Number of cores. Defines the number of cores of the molecule.
- **R:** For taking into account individual aliphatic chains and bridges between cores, this attribute requires some changes. For multicore molecules, it assumes a code of ten digits. The first two digits of each core represents the carbon number of its own aliphatic side chain. The remaining eight digits represent the carbon number of each possible core connection.

In Figure 4.26, we show an example of a multicore molecule represented by the structure-oriented lumping methodology.



A6	A4	A2	N6	N5	N4	N3	N2	N1	R	br	me	IH	AA	NS	RS	AN	NN	RN	NO	RO	KO	Cn	Nc
1	3	1	0	0	2	0	0	1	1605000000	1	3	0	1	2	0	0	1	0	0	0	0	20001000	2
1	2	0	0	0	0	0	1	0	1405000000	1	1	0	0	0	0	1	0	0	0	1	0	10001000	0

Figure 4.26: Example of a multicore molecule and its representation by the structure-oriented lumping vector

QUANN and JAFFE (1992), designed a stoichiometric matrix to be used in conjuction with the structure vectors. This matrix defines the total number of each atom type for each of the structural attributes. In Figure 4.27, we show the proposed matrix.

	A6	**A4**	**A2**	**N6**	**N5**	**N4**	**N3**	**N2**	**N1**	**R**	**br**	**me**	**IH**	**AA**	**NS**	**RS**	**AN**	**NN**	**RN**	**NO**	**RO**	**KO**
Carbon	6	4	2	6	5	4	3	2	1	1	0	0	0	0	−1	0	−1	−1	0	−1	0	0
Hydrogen	6	2	0	12	10	6	4	2	1	2	0	0	2	−2	−2	0	−1	−1	1	−2	0	0
Sulfur	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
Nitrogen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
Oxygen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1

Figure 4.27: Stoichiometry matrix for the structure-oriented lumping method (QUANN and JAFFE, 1992)

Using the matrix shown in Figure 4.27, one can calculate the total number of atoms for each molecule.:

$$\mathbf{MA} = \mathbf{M} \times \mathbf{S}^{\intercal} \tag{4.11}$$

Where $\mathbf{M}$ is the molecular matrix based on the structural attributes in Figure 4.25, and $\mathbf{S}$ is the stoichiometry matrix shown in Figure 4.27.

The resulting matrix, $\mathbf{MA}$, contains the number of rows corresponding to the number of molecules, and five columns representing each atom type. Each element of the matrix represents the total number of an atom type of the corresponding molecule. Since we are interested in the total number of atoms, when dealing with multicore molecules one can sum all the structural attributes of each core and put that information in a single row.

Extension of the structure-oriented lumping vector

Inspired by the stoichiometry matrix shown in Figure 4.27, we designed a functional group matrix related to the structure-oriented lumping vectors. However, in order to work with many functional groups as possible, an extension of the original vector proposed by QUANN and JAFFE (1992) was necessary. In Figure 4.28, we show the proposed extended structure-oriented lumping vector.

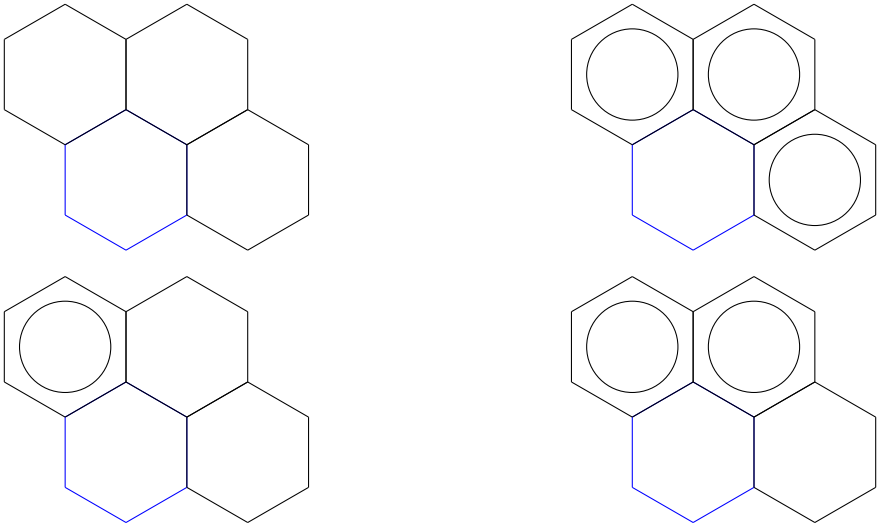
$$\begin{bmatrix} \textbf{A6} & \textbf{A4} & \textbf{A2} & \textbf{N6} & \textbf{N5} & \textbf{N4a} & \textbf{N4b} & \textbf{N3} & \textbf{N2n} & \textbf{N2a} & \textbf{N2m} & \textbf{N2f} & \textbf{N1a} & \textbf{N1m} & \textbf{N1n} & \textbf{Rp} & \textbf{Rm} & \textbf{Rn} \\ \textbf{Ra} & \textbf{br} & \textbf{me1} & \textbf{me2} & \textbf{IHo} & \textbf{IHn} & \textbf{AA1} & \textbf{AA2} & \textbf{AA3} & \textbf{NS} & \textbf{RS} & \textbf{AN} & \textbf{NN} & \textbf{RN} & \textbf{NO} & \textbf{RO} & \textbf{KO} & \end{bmatrix}$$

Figure 4.28: Extended version of the structure-oriented lumping vector

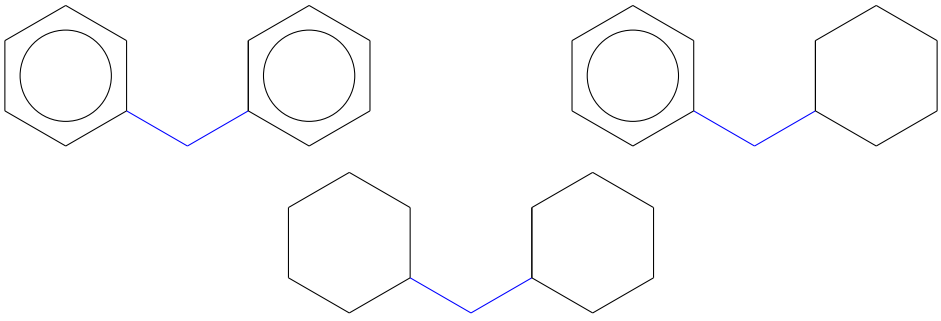
The additional attributes proposed are defined as follows:

- **N4a and N4n:** A four carbon incremental naphthenic ring connected to aromatic ring and a naphthenic ring, respectively.

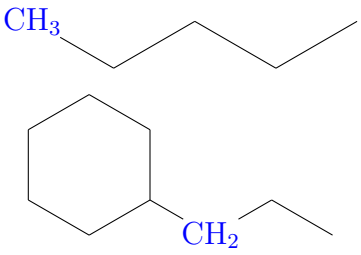
- **N2n, N2a, N2m, and N2f:** A two carbon incremental naphthenic ring connected to three naphthenic rings, to three aromatic rings, to two naphthenics and one aromatic rings, and to one naphthenic and two aromatic rings, respectively.

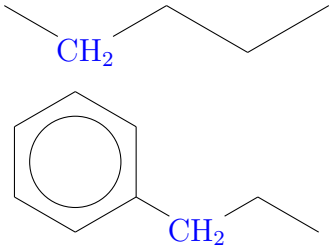


- **N1a, N1m, and N1n:** One carbon incremental naphthenic ring connected to two aromatic rings, one aromatic and one naphthenic ring and two naphthenic rings, respectively.

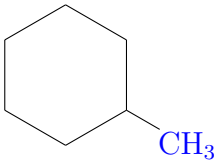


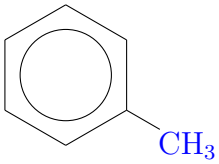
- **Rp, Rm, Rn, and Ra:** An aliphatic terminal carbon, an aliphatic carbon in the middle of a chain, an aliphatic carbon connected to a naphthenic ring, and an aliphatic carbon connected to an aromatic ring, respectively.



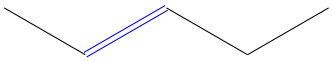


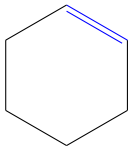
- **me1, and me2:** A methyl substitution to a naphthenic, and an aromatic ring, respectively.



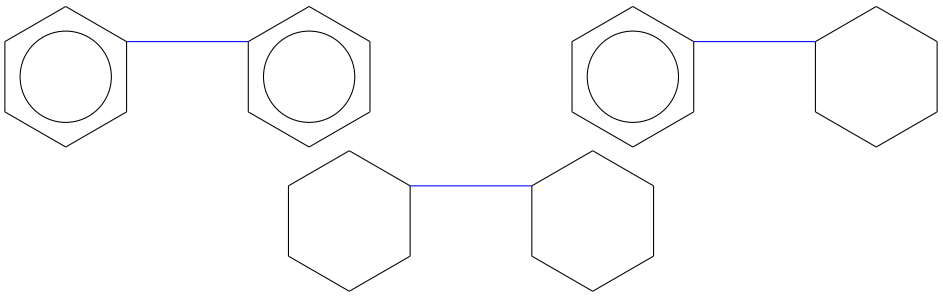


- **IHo, and IHn:** Degree of unsaturation in an aliphatic chain, and in a naphthenic ring, respectively.



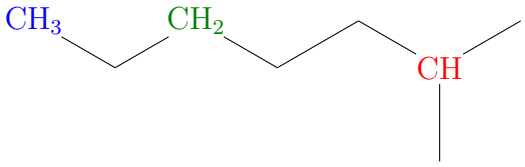


- **AA1, AA2, and AA3:** A biphenyl bridge connecting two aromatic rings, one aromatic ring and one naphthenic ring, and two naphthenic rings, respectively.

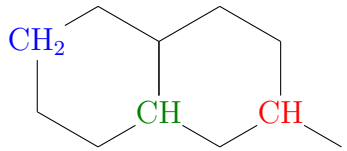


With these new attributes, we were able to design a functional groups matrix similar to the stoichiometry matrix presented in Figure 4.27. In this work, we considered the following functional groups:

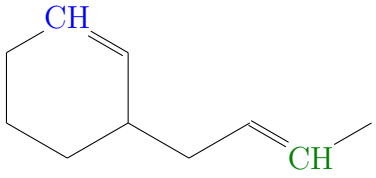
- $\mathbf{-CH_3-nr}$, $\mathbf{-CH_2-nr}$, and $\mathbf{>CH-nr}$: Aliphatic primary, secondary and tertiary carbons, respectively.



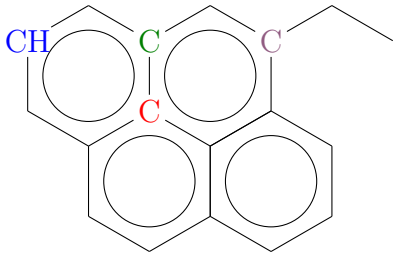
- $> \mathbf{CH_2} - \mathbf{r}$, $> \mathbf{CH} - \mathbf{r_1}$, and $> \mathbf{CH} - \mathbf{r_2}$: Naphthenic carbon, condensed naphthenic carbon and substituted naphthenic carbon, respectively.



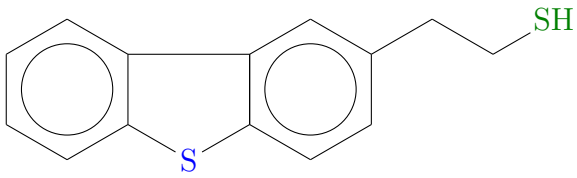
- $= \mathbf{CH} - \mathbf{nr}$, and $= \mathbf{CH} - \mathbf{r}$: Aliphatic olefinic carbons and cyclic olefinic carbons, respectively.



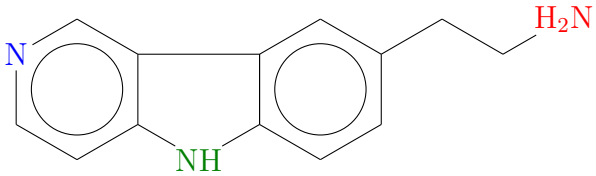
- $\mathbf{CH}$, $\mathbf{C_1}$, $\mathbf{C_2}$, and $= \mathbf{C} < \mathbf{r}$: Aromatic carbon, peripheral condensed aromatic carbon, internal condensed aromatic carbon, and substituted aromatic carbon, respectively.



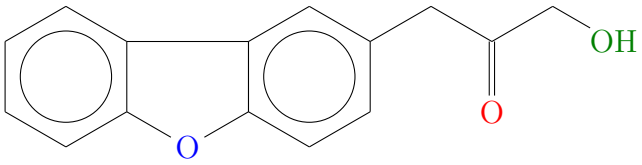
- $-\mathbf{S} - \mathbf{r}$, and $\mathbf{SH} - \mathbf{nr}$: Tiophenic sulfur, and aliphatic mercaptan sulfur, respectively.



- $= \mathbf{N} - \mathbf{r}$, $-\mathbf{NH} - \mathbf{r}$ and $-\mathbf{NH_2} - \mathbf{nr}$: Pyridinic nitrogen, pyrrolic nitrogen, and aliphatic amine nitrogen, respectively.



- $\mathbf{-O-r}$, $\mathbf{OH}$ and $\mathbf{>C=O-nr}$: Furanic oxygen, aliphatic alcohol oxygen, and aliphatic carbonyl oxygen, respectively.

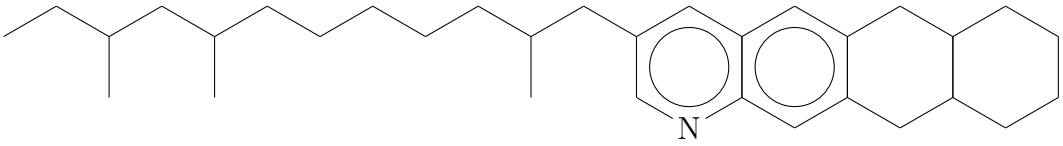


In Figure 4.30, we show the proposed matrix. It has the functional groups on the rows and the structural attributes on the columns. With this matrix, one can calculate the quantities of each functional group in each of the molecules:

$$\mathbf{MF} = \mathbf{M_{ex}} \times \mathbf{F}^{\intercal} \tag{4.12}$$

Where $\mathbf{M_{ex}}$ is the molecular matrix based on the extended structural attributes in Figure 4.28, and $\mathbf{F}$ is the functional groups matrix shown in Figure 4.30.

The resulting matrix, $\mathbf{MF}$, has the molecules in the rows and the functional groups in the columns. Each element contains the quantities of the corresponding functional group in the corresponding molecule. In Figure 4.29, we show an example molecule and its representation by the matrices $\mathbf{M}$, $\mathbf{M_{ex}}$, $\mathbf{MA}$, and $\mathbf{MF}$.



$$\mathbf{M} = \begin{bmatrix} \mathbf{A6} & \mathbf{A4} & \mathbf{A2} & \mathbf{N6} & \mathbf{N5} & \mathbf{N4} & \mathbf{N3} & \mathbf{N2} & \mathbf{N1} & \mathbf{R} & \mathbf{br} & \mathbf{me} & \mathbf{IH} & \mathbf{AA} & \mathbf{NS} & \mathbf{RS} & \mathbf{AN} & \mathbf{NN} & \mathbf{RN} & \mathbf{NO} & \mathbf{RO} & \mathbf{KO} \\ 1 & 1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 15 & 3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{M_{ex}} = \begin{bmatrix} \mathbf{A6} & \mathbf{A4} & \mathbf{A2} & \mathbf{N6} & \mathbf{N5} & \mathbf{N4a} & \mathbf{N4b} & \mathbf{N3} & \mathbf{N2n} & \mathbf{N2a} & \mathbf{N2m} & \mathbf{N2f} & \mathbf{N1a} & \mathbf{N1m} & \mathbf{N1n} & \mathbf{Rp} & \mathbf{Rm} & \mathbf{Rn} \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 14 & 0 \\ \mathbf{Ra} & \mathbf{br} & \mathbf{me1} & \mathbf{me2} & \mathbf{IHo} & \mathbf{IHn} & \mathbf{AA1} & \mathbf{AA2} & \mathbf{AA3} & \mathbf{NS} & \mathbf{RS} & \mathbf{AN} & \mathbf{NN} & \mathbf{RN} & \mathbf{NO} & \mathbf{RO} & \mathbf{KO} & \\ 1 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \end{bmatrix}$$

$$\mathbf{MA} = \begin{bmatrix} \mathbf{Carbon} & \mathbf{Hydrogen} & \mathbf{Sulfur} & \mathbf{Nitrogen} & \mathbf{Oxygen} \\ 36 & 51 & 0 & 1 & 0 \end{bmatrix}$$

$$\mathbf{MF} = \begin{bmatrix} \mathbf{-CH_3-nr} & \mathbf{-CH_2-nr} & \mathbf{>CH_2-r} & \mathbf{>CH-nr} & \mathbf{>CH-r_1} & \mathbf{>CH-r_2} & \mathbf{=CH-nr} & \mathbf{=CH-r} & \mathbf{CH} & \mathbf{C_1} \\ 4 & 8 & 6 & 3 & 2 & 0 & 0 & 0 & 4 & 4 \\ \mathbf{C_2} & \mathbf{=C<} & \mathbf{-S-r} & \mathbf{SH-nr} & \mathbf{=N-r} & \mathbf{-NH-r} & \mathbf{-NH_2-nr} & \mathbf{-O-r} & \mathbf{-OH} & \mathbf{>C=O-nr} \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 4.29: An example molecule and its matrix representation: Structure-oriented lumping, extended structure-oriented lumping, stoichiometry and functional groups.

	A6	A4	A2	N6	N5	N4a	N4b	N3	N2n	N2a	N2m	N2f	N1a	N1m	N1n	Rp	Rm	Rn
$-CH_3-nr$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
$-CH_2-nr$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
$>CH_2-r$	0	0	0	6	5	4	2	1	0	2	1	1	1	0	−1	0	0	−1
$>CH-nr$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$>CH-r_1$	0	0	0	0	0	0	2	2	2	0	1	1	0	1	2	0	0	0
$>CH-r_2$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
$=CH-nr$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$=CH-r$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CH	6	2	0	0	0	−2	0	0	0	−2	−1	−1	−2	−1	0	0	0	0
C_1	0	2	0	0	0	2	0	0	0	0	0	−1	2	1	0	0	0	0
C_2	0	0	2	0	0	0	0	0	0	2	1	2	0	0	0	0	0	0
$=C<$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$-S-r$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$SH-nr$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$=N-r$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$-NH-r$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$-NH_2-nr$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$-O-r$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$-OH$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$>C=O-nr$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(a)

	Ra	br	me1	me2	IHo	IHn	AA1	AA2	AA3	NS	RS	NA	NN	RN	NO	RO	KO
$-CH_3-nr$	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
$-CH_2-nr$	0	−2	−1	−1	−2	0	0	0	0	0	0	0	0	0	0	0	−1
$>CH_2-r$	0	0	−1	0	0	0	0	−1	−2	−1	0	0	−1	0	−1	0	0
$>CH-nr$	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$>CH-r_1$	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0
$>CH-r_2$	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$=CH-nr$	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
$=CH-r$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CH	−1	0	0	−1	0	0	−2	−1	0	0	0	−1	0	0	0	0	0
C_1	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0
C_2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$=C<$	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
$-S-r$	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
$SH-nr$	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
$=N-r$	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
$-NH-r$	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
$-NH_2-nr$	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
$-O-r$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
$-OH$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
$>C=O-nr$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

(b)

Figure 4.30: Functional groups versus structural attributes matrix. (a) columns 1 to 18. (b) columns 19 to 35

4.5 Properties calculation

In Section 4.3, we described the sampling methodology used in the stochastic reconstruction algorithm. Then, in Section 4.4, we presented a methodology to storage the probability density functions outputs in terms of molecular structures using the structure-oriented lumping vector created by QUANN and JAFFE (1992). We also proposed an extension of the structure-oriented lumping vector. After sampling from the probability density functions, the built molecules are represented in terms of four different matrices: **M**, **MA**, $\mathbf{M_{ex}}$, and **MF** (refer to Section 4.4).

In this section, we describe the methods used to calculate properties both for the individual molecules and for the hypothetical mixture. The building diagrams presented in Section 4.2 are sampled N times generating N molecules. In that sense, matrices **M**, **MA**, $\mathbf{M_{ex}}$, and **MF** also have N rows. At this point, we consider an equimolar mixture based on the N molecules created.

By combining the calculated properties with the experimental data, we are able to calculate likelihood functions (or at least an approximation of it). These likelihoods are used to estimate the probability density function parameters controlling the sampling step.

Pure molecules properties

Each molecule has its own set of physical and chemical properties. In this work, we used a basic set of properties to represent each molecule and its contribution to the mixture.

- **Molecular mass**. In Section 4.4, we defined the matrix **MA**. This matrix has molecules in each row and quantities of each atom type in each column. Defining **VMW** as the matrix of each atom type molecular mass:

$$\mathbf{VMW} = \begin{bmatrix} \textbf{Carbon} & \textbf{Hydrogen} & \textbf{Sulfur} & \textbf{Nitrogen} & \textbf{Oxygen} \\ 12 & 1 & 32 & 14 & 16 \end{bmatrix}$$

  The multiplication of vector **VMW** by the transpose of matrix **MA**, results in a vector of molecular masses, **MMW**, as follows:

$$\mathbf{MMW} = \mathbf{VMW} \times \mathbf{MA}^{\intercal} \tag{4.13}$$

- **Specific gravity and normal boiling points**. In order to calculte specific gravity and normal boiling points, we a used the group contribution method

described by DE OLIVEIRA *et al.* (2013). The normal boiling point is calculated as follows:

$$\exp\left(\frac{T_b}{307.63}\right) = \sum_i (n_i \Delta T_{b,i}) + F_{T_b} \tag{4.14}$$

$\Delta T_{b,i}$ is the contribution of group i and n_i is the number of groups of type i. F_{T_b} is a correction term based on the total number of rings, N_R, calculated as follows:

$$F_{T_b} = 0.2285 N_R^2 + 0.4678 N_R \tag{4.15}$$

For the specific gravity, which considers the molecule in a liquid state at $20\,^{\circ}C$, we used the following equation:

$$d = \frac{MW}{\sum_i n_i \Delta V_{m,i} + F_{V_m}} \tag{4.16}$$

$\Delta V_{m,i}$ is the contribution of group i and n_i is the number of groups of type i. MW is the molecular mass of the molecule and F_{V_m} is a correction term based on the total number of rings, N_R, calculated as follows:

$$F_{V_m} = 25 N_R \tag{4.17}$$

In Figure 4.31, we show the matrix **FGP**. This matrix contains the values of $\Delta T_{b,i}$ and $\Delta V_{m,i}$ for the functional groups considered in this work (refer to Figure 4.29). The terms $\sum_i n_i \Delta T_{b,i}$ and $\sum_i n_i \Delta V_{m,i}$ from Equations 4.14 and 4.16 can be easily calculted by:

$$\mathbf{MPM} = \mathbf{MF} \times \mathbf{FGP} \tag{4.18}$$

The resulting matrix, **MPM**, has the molecules in the rows and two columns. The columns represent the molecule structural contribution to the boiling point and specific gravity, respectively. N_R can be calculated for each molecule i by:

$$N_{R_i} = \sum_{j=1}^{j=9} \mathbf{M_{i,j}} \tag{4.19}$$

where $\mathbf{M}$ is the molecular matrix defined in Section 4.4. In matrix form, Equations 4.14 and 4.16 becomes, for each molecule i:

$$\exp\left(\frac{T_{b_i}}{307.63}\right) = \mathbf{MPM_{i,1}} + 0.2285\left(N_{R_i}\right)^2 + 0.4678 N_{R_i} \tag{4.20}$$

$$d_i = \frac{\mathbf{MMW_i}}{\mathbf{MPM_{i,2}} + 25 N_{R_i}} \tag{4.21}$$

$$\mathbf{FGP} = \begin{bmatrix} & \Delta T_{b,i} & \Delta V_{m,i} \\ \mathbf{-CH_3-nr} & 32.14 & 0.8758 \\ \mathbf{-CH_2-nr} & 16.38 & 0.3101 \\ \mathbf{>CH_2-r} & 13.93 & 0.3852 \\ \mathbf{>CH-nr} & -0.93 & -0.3343 \\ \mathbf{>CH-r_1} & -3.98 & -0.1343 \\ \mathbf{>CH-r_2} & -1.16 & -0.2519 \\ \mathbf{=CH-nr} & 13.55 & 0.3232 \\ \mathbf{=CH-r} & 10.97 & 0.3702 \\ \mathbf{CH} & 11.22 & 0.3814 \\ \mathbf{C_1} & -7.74 & 0.0067 \\ \mathbf{C_2} & -10.97 & -0.3995 \\ \mathbf{=C<} & -6.15 & -0.1494 \\ \mathbf{-S-r} & 12.34 & 0.8741 \\ \mathbf{SH-nr} & 12.53 & 0.9281 \\ \mathbf{=N-r} & -2.30 & 0.6455 \\ \mathbf{-NH-r} & -1.12 & 1.1069 \\ \mathbf{-NH_2-nr} & 8.58 & 0.5954 \\ \mathbf{-O-r} & 14.63 & 0.2540 \\ \mathbf{-OH} & 5.61 & 0.3232 \\ \mathbf{>C=O-nr} & 10.09 & 1.0150 \end{bmatrix}$$

Figure 4.31: Functional groups contributions to specific gravity and normal boiling points calculation.

Mixture properties

In order to estimate the properties for the hypothetical mixture, we need to make some assumptions. First, we considered that the mixture is ideal. This means that we excluded the effect of molecular interactions in the calculations. This is specifically important to the average molecular mass, specific gravity and boiling point curve. Second, the same importance is given to every molecule built from the building diagrams presented in Section 4.2. In that sense, every molecule has the same mole fraction given by:

$$x_i = \frac{1}{N} \tag{4.22}$$

where N is the total number of molecules.

- **Average molecular mass:** The average molecular mass, MW_{avg}, is simply the sum of the individual molecular masses weighted by the mole fraction. Defining $\mathbf{X}$ as the mole fraction vector:

$$MW_{avg} = \mathbf{MMW} \times \mathbf{X} \tag{4.23}$$

- **Mass fractions:** After calculating the average molecular mass, one can calculate the mass fraction of each molecule, w_i, by:

$$w_i = \frac{\mathbf{X_i} \times \mathbf{MMW_i}}{MW_{avg}} \tag{4.24}$$

- **Mixture specific gravity:** Considering an ideal mixture, the inverse of the specific gravity is additive in a mass basis. The mixture specific gravity, $S_{g_{mix}}$, is simply the sum of the inverse of the individual specific gravities weighted by the mass fraction:

$$S_{g_{mix}} = 1/\sum_i \frac{w_i}{d_i} \tag{4.25}$$

- **Volume fractions:** After calculating the mixture specific gravity, one can calculate the volume fraction of each molecule, v_i, by:

$$v_i = \frac{w_i S_{g_{mix}}}{d_i} \tag{4.26}$$

- **Boiling point curve:** This analysis reports the boiling point associated with a certain amount of sample vaporization. This amount of vaporization can be reported in a mass or volumetric basis. For instance, some vacuum residue can have a 10% vaporization temperature of $550\,°C$. We used the following procedure to calculate this property:

  - First we arrange the molecules in a crescent order of boiling points.

  - Then, we calculate the cumulative mass (cw_i) or volume (cv_i) percentages of each molecule. For example, consider that molecule 1 has a mass fraction of 5% and molecule 2 has a mass fraction of 8%. The cumulative mass percentage for molecule 2 is 13%.

  - Finally, to calculate the boiling point associated with a k% vaporization, we look for the molecules with a cumulative mass fraction immediately below and above k%. Then, the boiling point is a linear interpolation between the boiling points of the two individual molecules:

$$T_{b,k\%} = T_{b_{i<k\%}} + \frac{(k\% - cw_{i<k\%})(T_{b_{i>k\%}} - T_{b_{i<k\%}})}{cw_{i>k\%} - cw_{i<k\%}} \tag{4.27}$$

- **SARA fractions:** This method separates petroleum fractions into four groups: saturates, aromatics, resins, and asphaltenes. The results are just the mass factions of each class. However, because separation based on solubility is hard to classify, here, we used the criteria proposed by DE OLIVEIRA *et al.* (2013), which separates the molecules based on the mass percentage of hydrogen and molecular mass. The criteria are as follows:

$$\text{if } \%H_i \leq 14 - \frac{11300}{\mathbf{MMW_i} + 800} = \text{Asphaltene} \tag{4.28}$$

$$\text{if } \%H_i \leq 14 - \frac{4000}{\mathbf{MMW_i} + 160} = \text{Resin} \tag{4.29}$$

$$\text{if } \%H_i \leq 14 - \frac{3000}{\mathbf{MMW_i} + 1300} = \text{aromatics} \tag{4.30}$$

the molecule hydrogen mass percentage can be easily calculated by:

$$\%H_i = 100 \times \frac{\mathbf{MA_{i,2}} \times \mathbf{VMW_{1,2}}}{\sum_j \mathbf{MA_{i,j}} \times \mathbf{VMW_{1,j}}} \tag{4.31}$$

- **Elemental analysis:** This analysis measures the mass percentage of each of the main atom types in the mixture (Carbon, Hydrogen, Sulfur, Nitrogen, and Oxygen). The mass percentage of an atom type k in the mixture is the sum of each molecule mass percentage of the same atom weighted by the mass fraction

$$w_{k_{mix}} = \sum_{i=1}^{N} \left(\frac{x_i \times \mathbf{MA_{i,k}} \times \mathbf{VMW_{1,k}}}{\sum_{i=1}^{N} x_i \times \mathbf{MMW_i}} \right), \tag{4.32}$$

- **Nuclear magnetic resonance:** Similar to the elemental analysis, this method measures the molar percentage of certain carbon (or hydrogen) type relative to the total quantity of carbons (or hydrogen). This atom types are described in Section 4.1, and are directly related to the functional groups in matrix **MF** (refer to Section 4.4). The molar percentage of certain atom type k is then

$$x_{k_{mix}} = \left(\frac{\sum_i \mathbf{MF_{i,k}} \times \mathbf{X_i}}{\sum_i \mathbf{MA_{i,k}} \times \mathbf{X_i}} \right) \tag{4.33}$$

In this chapter, we described the stochastic reconstruction algorithm developed in this thesis. Using the experimental data available, we defined the chemical attributes to be modeled by probability density functions. The sampling protocol, functional forms and parameters of the probability density functions were defined based on prior chemical knowledge, experimental evidence and available data. We finished the chapter describing the methodology for molecular representation and properties calculation. The probability density functions parameters control the molecular generation properties. In that sense, they should be estimated to better match the experimental data available. We can define the stochastic reconstruction method as a generative model for simulated data. This simulated data can then be used in a parameter inference methodology, as described in Chapter 5.

Chapter 5

Statistical inference

In this chapter, we describe the methodology used to estimate the parameters of the molecular reconstruction algorithm. We also describe the clustering technique used to select the best molecules candidates and composition calculation by entropy maximization. As described in Chapter 4, the molecular reconstruction algorithm can be seen as a parametrized stochastic data generating mechanism. DUTTA *et al.* (2016) defined this type of data generating process as simulator-based models. In practical terms, it is a computer program that takes a value θ and a state of the random number generator as input and returns data y_θ as output (GUTMANN and CORANDER, 2016).

5.1 The likelihood principle

In statistical inference, one uses the information in a data sample $Y_1, \ldots, Y_m$ to make inferences about an unknown parameter θ. In this thesis, we used the likelihood principle as the data reduction device to find suitable estimators for the unknown parameters. This principle is based on an important statistic called the likelihood function.

Let $p(y|\theta)$ denote the joint probability density function of a data sample $Y = (Y_1, \ldots Y_n)$. Then, given that $Y = y$ is observed, the function of θ defined by

$$L(\theta|y) = p(y|\theta) \tag{5.1}$$

is called the likelihood function (CASELLA and BERGER, 2002).

If Y is a discrete random variable, then $L(\theta|y) = P_\theta(Y = y)$. If we compare the

likelihood function at two parameter points and find that

$$P_{\theta_1}(Y = y) = L(\theta_1|y) > P_{\theta_2}(Y = y) = L(\theta_2|y), \tag{5.2}$$

then the sample we actually observed is more likely to have occurred if $\theta = \theta_1$ then if $\theta = \theta_2$, which can be interpreted as saying that θ_1 is a more plausible value for the true value of θ than is θ_2 (CASELLA and BERGER, 2002).

The likelihood principle can then be defined as follows: if x and y are two sample points such that $L(\theta|x)$ is proportional to $L(\theta|y)$, that is, there exists a constant $C(x, y)$ such that

$$L(\theta|x) = C(x, y)L(\theta|y), \tag{5.3}$$

then the conclusions drawn from x and y should be identical (CASELLA and BERGER, 2002).

Maximum likelihood estimators

The likelihood principle and the interpretation of likelihood values led to one of the most popular techniques for deriving estimators. If $Y_1, \ldots, Y_n$ are an independent and identically distributed sample from a population with probability density function $f(y|\theta_1, \ldots, \theta_k)$, the likelihood function is defined by (CASELLA and BERGER, 2002)

$$L(\theta|y) = L(\theta_1, \ldots, \theta_k|y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i|\theta_1, \ldots, \theta_k). \tag{5.4}$$

For each sample point y, let $\hat{\theta}(y)$ be a parameter value at which $L(\theta|y)$ attains its maximum as a function of θ, with y held fixed. A maximum likelihood estimator of the parameter θ based on a sample Y is $\hat{\theta}(Y)$ (CASELLA and BERGER, 2002).

Bayes estimators

In statistical inference there exist fundamentally different approaches regarding the estimation of unknown parameters. In the classical approach the parameter, θ, is thought to be an unknown, but fixed, quantity (CASELLA and BERGER, 2002). By fixed, we mean that it can not be considered a random variable. A random sample $Y_1, \ldots, Y_n$ is drawn from a population indexed by θ and, based on the observed values in the sample, knowledge about the value θ is obtained (CASELLA and BERGER, 2002). For that, one can use the likelihood principle and maximum likelihood estimators.

In the Bayesian approach, θ is considered to be a quantity whose variation can be described by a probability distribution, called the prior distribution. This distribution is inherently subjective, since it is based on the experimenters' belief and defined before any data is observed. In a similar way, after some random sample from a distribution indexed by θ is observed, one can update the prior distribution of θ using Bayes rule. The updated distribution is called the posterior distribution.

If we denote the prior distribution by $\pi(\theta)$ and the sampling distribution by $f(y|\theta)$, then the posterior distribution, the conditional distribution of θ given the sample, y, is

$$\pi(\theta|y) = f(y|\theta)\pi(\theta)/m(y), \tag{5.5}$$

where $m(y)$ is the marginal distribution of Y, that is,

$$m(x) = \int f(y|\theta)\pi(\theta)d\theta. \tag{5.6}$$

The likelihood principle is a valid data reduction device for either one of the approaches, and as shown in Equation 5.5, one uses the likelihood function to update the prior distributions. The posterior distribution contains all the information regarding the parameter θ. The mean of this distribution can be considered a point estimator for θ.

In a nutshell, classical methods of inference portrait the parameter as a fixed and unknown quantity and the observed data as random. In contrast, Bayesian methods portrait the parameter as the random quantity and the observed data as fixed. In some sense, classical methods draw conclusions about what might have been observed and Bayesian methods about what was actually observed.

In this work, we chose the Bayesian approach to statistical inference. As we will discuss later, our model has an intractable likelihood function. For such complex models, the Bayesian approach is more suitable, especially for estimating parameter uncertainty.

5.2 Approximate Bayesian computation

In Chapter 4, we described the stochastic reconstruction algorithm. It models chemical attributes using probability density functions. These functions have parameters that control their shape and, in turn, the data generated by the model. The data generation process follows a Monte Carlo type procedure, by taking random samples

from the probability density functions. The raw data generated by the model are the chemical attributes for each sample, which can be translated into molecules. In that sense, stochastic reconstruction algorithms are functions M that map the model parameters θ and some random variables V to data y (DUTTA *et al.*, 2016).

The presence of the random variables V causes the outputs of the model to randomly fluctuate even when the parameters θ are held fixed. To discuss the properties of simulator-based models, such as the stochastic reconstruction algorithm, we will follow the definitions described in DUTTA *et al.* (2016). Due to the random fluctuations caused by the variables V the simulator defines a random variable Y_θ with a distribution implicitly determined by the distribution of V for a given θ.

As described in DUTTA *et al.* (2016), for a fixed value of θ, the probability that Y_θ takes values in an ϵ neighborhood $B_\epsilon(y_0)$ around the observed data y_0 is equal to the probability to draw values of V that are mapped to that neighborhood,

$$Pr(Y_\theta \in B_\epsilon(y_0)) = Pr(M(\theta, V) \in B_\epsilon(y_0)). \tag{5.7}$$

Computing this probability analytically is impossible for complex models (DUTTA *et al.*, 2016). However, it is clear that one can obtain some particular data from the simulator, y_θ, and test if it ends up in the neighborhood of the observed data, y_0. Equation 5.7 can be used to define a likelihood function as ϵ approaches zero,

$$L(\theta|y_0) = \lim_{\epsilon \to 0} c_\epsilon Pr(Y_\theta \in B_\epsilon(y_0)). \tag{5.8}$$

However, neither the probability in Equation 5.7 nor the likelihood function in Equation 5.8 are available analytically. Considering Y_θ a discrete random variable, the likelihood function can be written as

$$L(\theta) = Pr(Y_\theta = y_0). \tag{5.9}$$

From Equation 5.9, one can see that for discrete random variables the likelihood is either 0 or 1. For simulator-based models, even with intractable likelihoods, one can simulate from the model and check if the generated data matches the observed data. In order to sample from the posterior distribution showed in Equation 5.5, the following algorithm can be used: samples are taken from the prior distribution of parameters $\theta \sim \pi(\theta)$ and the generated data are compared to the observed data. If $Y_\theta = y_0$, the samples are valid posterior samples. This algorithm is known as

the rejection sampling method and it returns an exact posterior distribution of parameters. The procedure is shown in Algorithm 5.1

Algorithm 5.1 Rejection sampling applied to simulator-based models to produce N independent samples from the posterior distribution

1: **for** i = 1 **to** N **do**
2: **repeat**
3: Generate θ from the prior $\pi(\theta)$
4: Generate y_θ from the model
5: **until** $y_\theta = y_0$
6: $\theta^{(i)} \leftarrow \theta$
7: **end for**

In practical terms, Y_θ can assume an uncountable number of values turning the probability in Equation 5.9 negligibly small. For instance, that is the case if Y_θ is a continuous random variable. In such cases, an already very inefficient algorithm becomes unfeasible.

To overcome this issue, one can replace the acceptance criteria in Algorithm 5.1 by

$$d(y_\theta, y_0) \leq \epsilon, \tag{5.10}$$

where $d(y_\theta, y_0)$ is a distance function and ϵ is a threshold defined by the modeler. The distance function is here to measure the discrepancy between simulated and experimental data and is usually applied to summary statistics of the data. For stochastic reconstruction algorithms, the use of summary statistics is a requirement. The model generates chemical attributes as simulated data. Clearly, one can only measure overall properties. By calculating overall properties, we are in fact calculating summary statistics of the generated data.

This new criteria leads to Algorithm 5.2, known as the approximate Bayesian computation rejection algorithm.

Algorithm 5.2 Approximate Bayesian computation rejection sampling

1: **for** i = 1 **to** N **do**
2: **repeat**
3: Generate θ from the prior $\pi(\theta)$
4: Generate y_θ from the model
5: **until** $d(y_\theta, y_0) \leq \epsilon$
6: $\theta^{(i)} \leftarrow \theta$
7: **end for**

Algorithm 5.2 does not generate samples from the posterior distribution in Equation 5.5. Instead, it produces an approximate posterior distribution conditional on $d(y_\theta, y_0) \leq \epsilon$. For that, it approximates the likelihood function $L(\theta)$ by $L_{d,\epsilon}(\theta)_{abc}$, as follows:

$$L_{d,\epsilon}(\theta)_{abc} \propto Pr(d(Y_\theta, y_0) \leq \epsilon). \tag{5.11}$$

Clearly, the approximation is highly dependent on the choice of the distance function, the summary statistics and the threshold. The choice of sufficient summary statistics should be ideal for these kinds of problems. However, this is not an easy task. For the purpose of this work, the choice of summary statistics is limited to the available experimental data. The threshold is a trade-off parameter. Higher its value worst the approximation. In contrast, as it goes lower, the computational cost rises rapidly, due to the low probabilities of matching the $d(y_\theta, y_0) \leq \epsilon$ criteria. This is even more important when running the simulator by itself is costly, as it is the case for stochastic reconstruction algorithms.

Different algorithms have been designed to improve the rejection sampling algorithm efficiency. The most popular examples are based on Markov Chain Monte Carlo Methods (MARJORAM *et al.*, 2003) and the sequential Monte Carlo framework (BEAUMONT *et al.*, 2002). In a nutshell, these methods avoid parameters propositions drawn directly from the prior distribution, by iteratively morphing the prior into the posterior with some added noise. However, the rejection step is still present. In order to effectively sample from the posterior, millions of simulations are necessary especially for low values of the threshold.

For computationally costly simulators, such as the stochastic reconstruction algorithm, the required number of simulations turns inference unfeasible. In this work, we used the proposition from GUTMANN and CORANDER (2016) for improving computational efficiency. The methodology is called Bayesian optimization for likelihood-free inference and it is described in Section 5.3

5.3 Bayesian optimization for likelihood-free inference

In Section 5.2, we discussed the basic fundamentals of approximate Bayesian computation algorithms. These methods are used to estimate posterior distributions of parameters for models with intractable likelihoods. They are based on the estimation of a computable likelihood function, hence the approximation. For that, most methods rely on a rejection step: $d(y_\theta, y_0) \leq \epsilon$. The rejection step is the main

cause of the high computational cost. We are usually interested in regions where the discrepancy $d(y_\theta, y_0)$ is small. Those regions have a small acceptance probability, requiring millions of simulations to accurately estimate the posterior distribution. Add to that a computationally costly simulator and inference becomes unfeasible.

In GUTMANN and CORANDER (2016), a likelihood approximation based on regression is proposed. A probabilistic model relating the discrepancy $d(y_\theta, y_0)$ to the parameters θ is built. Bayesian optimization is then used to actively choose the training set of the model. Once trained, we can use this model as an approximation of the likelihood function, as we will see later. Two main improvements in computational efficiency are clear at this point. First, we are actively searching in regions of interest. Second, once the model is trained, no further simulations from the simulator are necessary.

Bayesian optimization and Gaussian processes

Bayesian optimization can be regarded as a method to finding the extrema of black-box functions(BROCHU *et al.*; GUTMANN and CORANDER, 2010; 2016). By black-box, we mean functions of unknown analitycal form and derivatives. That is the case for the simulator-based model described in this thesis. We assume that the discrepancy can be modeled by a gaussian distribution,

$$d(y_\theta, y_0) \sim \mathcal{N}(f(\theta), \sigma_n^2). \tag{5.12}$$

Once again, we rely on the fact that we can generate values of the discrepancy $d(y_\theta, y_0)$ for a given set of parameters θ from the simulator.

In Bayesian optimization, we use Bayes rule to infer the posterior distribution over possible functions given the observed data. Then, one chooses the next point to evaluate the simulator by optmization of an acquisition function over the posterior distribution of functions. The method can be divided in two steps: first a *surrogate* function is estimated based on evidence. The evidence is a set $\mathcal{E}_{1:t} = \{[d(y_{\theta_1}, y_0), \theta_1], \ldots [d(y_{\theta_t}, y_0), \theta_t]\}$ obtained from the simulator. Second, optimization of an acquisition function gives some points θ_{t+1} for gathering more evidence and update the *surrogate* function. This process continues untill convergence is attained.

The function $f(\theta)$ is often modeled as a Gaussian process (GUTMANN and CORANDER, 2016). In this work, we used this consideration. A Gaussian process is an extension of the multivariate Gaussian distribution to an infinite-dimension stochastic process for which any finite combination of dimensions will be a Gaussian distribution (BROCHU *et al.*, 2010). One can consider a Gaussian process as a

function, but instead of returning a scalar $f(\theta)$ for an arbitrary θ, it returns the mean and variance of a normal distribution over the possible values of f at θ (BROCHU *et al.*, 2010).

Placing a Gaussian process prior on $f(\theta)$

$$f(\theta) \sim \mathcal{GP}(\mu(\theta), k(\theta, \theta')). \tag{5.13}$$

We chose $\mu(\theta) = 0$, and as described in GUTMANN and CORANDER (2016), $k(\theta, \theta')$ was considered a squared exponential covariance function,

$$k(\theta, \theta') = \sigma_f^2 \exp\left(\sum_j \frac{1}{l_j^2}(\theta_j - \theta'_j)^2\right) \tag{5.14}$$

Given evidence $\mathcal{E}_{1:t} = \{[d(y_{\theta_1}, y_0), \theta_1], \ldots [d(y_{\theta_t}, y_0), \theta_t]\}$, the posterior probability density function of f at a point θ is Gaussian with posterir mean $m_{1:t}(\theta)$ and posterior variance $\nu_{1:t}^2(\theta)$ (JÄRVENPÄÄ *et al.*, 2019),

$$f(\theta)|\mathcal{E}_{1:t} \sim \mathcal{N}(m_{1:t}(\theta), \nu_{1:t}^2(\theta))), \tag{5.15}$$

where,

$$m_{1:t}(\theta) = k(\theta, \theta_{1:t})K(\theta_{1:t})^{-1}d(y_\theta, y_0)_{1:t}, \tag{5.16}$$

$$\nu_{1:t}^2(\theta) = k(\theta, \theta) - k(\theta, \theta_{1:t})K(\theta_{1:t})^{-1}k(\theta_{1:t}, \theta), \tag{5.17}$$

$$d(y_\theta, y_0)_{1:t} = (d(y_\theta, y_0)_1, \ldots \ d(y_\theta, y_0)_t)^\intercal, \tag{5.18}$$

$$k(\theta, \theta_{1:t}) = (k(\theta, \theta_1), \ldots, k(\theta, \theta_t))^\intercal, \tag{5.19}$$

$$K(\theta_{1:t}) = k(\theta_{1:t}, \theta_{1:t}) + \sigma_n^2\mathbf{I}, \tag{5.20}$$

$$k(\theta_{1:t}, \theta_{1:t})_{ij} = k(\theta_i, \theta_j) \quad \text{for} \quad i, j = 1, \ldots, t \tag{5.21}$$

At this point, we have a probabilistic model for the discrepancies $d(y_\theta, y_0)$ as a function of the parameters θ. Since the discrepancies are always positive, we modeled the logarithm of the discrepancies instead. With this model, we can approximate the likelihood function as follows (for more details refer to GUTMANN and CORANDER, 2016).

$$L_{d,\epsilon}(\theta)_{abc} \propto F\left(\frac{\log\epsilon - m_{1:t}(\theta)}{\sqrt{\nu^2_{1:t}(\theta)+\sigma^2_n}}\right), \tag{5.22}$$

where ϵ is the threshold, defined to be the $0.01th$ quantile of the realized discrepancies, and F is the cumulative distribution function of a standard normal random variable defined by

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt. \tag{5.23}$$

The last point to discuss in the Bayesian optimization framework is the optimization step itself. The optimization step is designed to actively choose the next point θ_{t+1} to be evaluated by the model and augment the evidence set. This is usually done by means of an acquisition function (BROCHU *et al.*; GUTMANN and CORANDER; JÄRVENPÄÄ *et al.*, 2010; 2016; 2019). Clearly, one has to set an objective for the optimizer. One possible choice is to search for regions with small discrepancies, as done in GUTMANN and CORANDER (2016).

JÄRVENPÄÄ *et al.* (2019) proposed efficient acquisition rules specifically designed for approximate Bayesian computation. Instead of looking for regions with small discrepancies, their method chooses the next evaluation point based on the expected uncertainty of the posterior distribution. Under the Gaussian process model, the point estimate for the expected value and variance of the unnormalized approximate posterior distribution $\pi^{ABC}(\theta|\mathcal{E}_{1:t})$ is given by (JÄRVENPÄÄ *et al.*, 2019)

$$\mathbb{E}(\pi^{ABC}(\theta|\mathcal{E}_{1:t})) = \pi(\theta)F\left(\frac{\log\epsilon - m_{1:t}(\theta)}{\sqrt{\nu^2_{1:t}(\theta)+\sigma^2_n}}\right), \tag{5.24}$$

$$\begin{aligned}\mathbb{V}(\pi^{ABC}(\theta|\mathcal{E}_{1:t})) = \pi^2(\theta)\Bigg[&F\left(\frac{\log\epsilon - m_{1:t}(\theta)}{\sqrt{\nu^2_{1:t}(\theta)+\sigma^2_n}}\right)F\left(\frac{\log\epsilon - m_{1:t}(\theta)}{\sqrt{\nu^2_{1:t}(\theta)+\sigma^2_n}}\right) \\ &-2T\left(\frac{\log\epsilon - m_{1:t}(\theta)}{\sqrt{\nu^2_{1:t}(\theta)+\sigma^2_n}}, \frac{\sigma_n}{\sqrt{2\nu^2_{1:t}(\theta)+\sigma^2_n}}\right)\Bigg],\end{aligned} \tag{5.25}$$

where $T(h,a)$ is Owen's t-function defined as

$$T(h,a) = \frac{1}{2\pi}\int_0^a \frac{e^{-h^2(1+x^2)/2}}{1+x^2}dx. \tag{5.26}$$

We then choose the next point of evaluation θ_{t+1} by taking random samples from the variance surface $\pi_q(\theta) \propto \mathbb{V}(\pi^{ABC}(\theta|\mathcal{E}_{1:t}))$. In that sense, $\theta_{t+1} \sim \pi_q(\theta)$. For that, we used the methods described in Section 5.4.

The likelihood function in Equation 5.22 is tractable. More than that, it is cheap to evaluate since no further runs of the simulator are necessary. Conventional methods for Bayesian inference can be used as we will discuss in Section 5.4. In Algorithm 5.3, we summarize the Bayesian optimization procedure.

Algorithm 5.3 Bayesian optimization for likelihood-free inference algorithm. Estimation of the likelihood function based on a evidence set of N data points for the $\log d(y_\theta, y_0)$ as a function of θ

for i = 1 **to** t **do**
    Generate θ from the prior $\pi(\theta)$
    Generate initial evidence set $\mathcal{E}_{1:t}$ from the simulator
    Calculate the posterior distribution of the Gaussion process for the $\log d(y_\theta, y_0)$ as a function of θ
end for
for i = t **to** N **do**
    Find θ_i by sampling from the distribution $\pi_q(\theta) \propto \mathbb{V}(\pi^{ABC}(\theta|\mathcal{E}_{1:t}))$
    Generate new evidence $\mathcal{E}_{1:(t+i)}$ from the simulator
    Augment the evidence set and update the Gaussian process
end for

5.4 Markov chain simulation

In Section 5.3, we discussed the approximation of the likelihood function using regression. The regression function models the relation between the discrepancy and the parameters using a Gaussian process. We can then use the likelihood function in Equation 5.22 to infer the posterior distribution of θ using Bayes rule (Equation 5.5). The methods discussed so far are designed for intractable likelihood problems, however, the marginal distribution in Equation 5.5 is also impossible (or not computationally efficient) to compute. In this section, we discuss the methods used to effectively sample from the posterior distribution.

As described in GELMAN *et al.* (2014), Markov chain simulation is a general method based on drawing values of θ from approximate distributions and then correcting those draws to better approximate the posterior target distribution, $p(\theta|y)$. In practical terms, one just needs to calculate $p(\theta|y)$ up to a normalizing constant, avoiding the computation of the marginal distribution in Equation 5.5. The sampling is done sequentially, with the distribution of the sampled draws depending

only on the last value drawn; hence, the draws from a Markov chain (GELMAN *et al.*, 2014).

Metropolis algorithm

To illustrate the use of Markov chain simulation in Bayesian computation, we will describe the Metropolis algorithm. Our description is based on GELMAN *et al.* (2014). The procedure is shown in Algorithm 5.4.

Algorithm 5.4 Metropolis algorithm applied to Bayesian computation.

1: Generate a start point θ^0 from the prior $\pi(\theta)$
2: **for** t = 1 **to** T **do**
3: Sample a proposal θ^* from a proposal distribution $J_t(\theta^*|\theta^{t-1})$
4: Calculate the ratio of densities $r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}$
5: $\theta^t = \theta^*$ with probability $\min(r, 1)$
6: **end for**

The Algorithm 5.4 can be described as a random walk with an acceptance/rejection rule to converge to the specified target distribution. Since we are only concerned with densities ratios, we eliminate the normalizing constant from Equation 5.5. The random walk behavior of the Metropolis algorithm makes convergence very slow, especially for high dimensional cases. In this work, we used the Hamiltonian Monte Carlo algorithm to sample both from the variance surface and from the approximate posterior distribution.

Hamiltonian Monte Carlo

The main characteristic of the Hamiltonian Monte Carlo algorithm is the substitution of the random walk behavior for a movement pattern analogous to physical dynamics. For that, for each component θ_j in the parameter space, we add a momentum variable ϕ_j. We can look at this algorithm as a variant of the Metropolis algorithm, in which the proposal distribution for θ is basically determined by ϕ.

For that, Hamiltonian Monte Carlo augments the posterior density by an independent distribution $p(\phi)$ on the momenta, thus defining a joint distribution, $p(\theta, \phi|y) = p(\phi)p(\theta|y)$ (GELMAN *et al.*, 2014). The augmented posterior density acts like a sort of potential energy controlling the trajectories in the parameter space. In that sense, Hamiltonian Monte Carlo requires gradients of the posterior density. This gradient represents the forces acting in the momentum distribution.

It is usual to give ϕ a multivariate normal distribution with mean 0 and covariance set to a prespecified mass matrix M (GELMAN *et al.*, 2014). It is common

to consider the mass matrix diagonal. More than that, usually the identity matrix is used as the mass matrix. The steps of a typical Hamiltonian Monte Carlo are shown in Algorithm 5.5.

Algorithm 5.5 Hamiltonian Monte Carlo algorithm applied to Bayesian computation.

for t = 1 **to** T **do**
  Draw ϕ from its distribution $\phi \sim N(0, M)$.
  for l = 1 **to** L **do**
    $\phi \leftarrow \phi + \frac{1}{2}\varepsilon \frac{d \log p(\theta|y)}{d\theta}$
    $\theta \leftarrow \theta + \varepsilon M^{-1}\phi$
    $\phi \leftarrow \phi + \frac{1}{2}\varepsilon \frac{d \log p(\theta|y)}{d\theta}$
  end for
  At this point, we have momentum and parameter values at the start of the
  updating process ϕ^{t-1}, θ^{t-1} and after the updating process ϕ^*, θ^*.
  Calculate the ratio of densities $r = \frac{p(\theta^*|y)p(\phi*)}{p(\theta^{t-1}|y)p(\phi^{t-1})}$
  $\theta^t = \theta^*$ with probability $\min(r, 1)$
end for

In Algorithm 5.5, the proposal distribution follows a trajectory in the parameter space for L steps, where the position of the parameters θ and its momentum ϕ are updated by a step size of ε. One may recognize the updating phase as the leapfrog algorithm from physics dynamics (the reader should refer to BETANCOURT, 2017 for a conceptual discussion regarding the Hamiltonian Monte Carlo method). After that, an acceptance/rejection rule is used to define the new set of parameters. Two new parameters are defined: L representing the number of leapfrog steps, and ε representing the step size. In order to avoid hand-tunning these parameters, we used a variant of the Hamiltonian Monte Carlo Algorithm called No-U-Turn Sampler (HOFFMAN and GELMAN, 2014). As described in GELMAN *et al.* (2014), instead of running for a fixed number of steps, L, the trajectory in each iteration continues until it turns around.

Convergence

In order to access convergence of the Markov chain we used the following strategy: first, we simulate at least two different sequences with overdispersed starting points. This allows us to check if each sequence converges to the same values of the estimands (by estimands, we mean all the parameters in the model and any other quantities of interest). Second, we discard the first half of the simulations. This first half is usually called the warmup steps.

The approach used to diagnose convergence is by checking mixing and stationarity (GELMAN *et al.*, 2014). After discarding the first half of the simulations, we split the remaining points, from each sequence, into two different sequences. For example, if we simulate 3 chains we end up, after splitting, with 6 chains. As described by GELMAN *et al.* (2014), this allows us to simultaneously test mixing (if all the chains have mixed well, the separate parts of the different chains should also mix) and stationarity (at stationarity, the first and second half of each sequence should be traversing the same distribution).

Let m be the number of chains after splitting and n be the lenght of the respective chain. For each parameter θ, we label the simulations as $\theta_{ij}(i = 1, \ldots, n; j = 1, \ldots, m)$, and we compute B and W, the between- and within-sequence variances (GELMAN *et al.*, 2014)

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta_{.j}} - \bar{\theta_{..}})^2, \quad \text{where} \quad \bar{\theta_{.j}} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij}, \quad \bar{\theta_{..}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta_{.j}} \tag{5.27}$$

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \quad \text{where} \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_{.j})^2. \tag{5.28}$$

The marginal posterior variance of the parameter can then be estimated by

$$\text{var}(\theta|y) = \frac{n-1}{n}W + \frac{1}{n}B. \tag{5.29}$$

To monitor convergence, one can calculate a potential scale reduction, which declines to 1 as $n \to \infty$

$$\hat{R} = \sqrt{\frac{\text{var}(\theta|y)}{W}}. \tag{5.30}$$

If the potential scale reduction is high, then we have reason to believe that proceeding with further simulations may improve our inference about the target distribution of the associated parameter (GELMAN *et al.*, 2014).

Another important diagnosis parameter for Markov chain simulations is the number of effective samples. In some sense, it measures the amount of correlation between individual samples. Higher values indicate a low correlation between samples, which is something we are looking for. The number of effective samples can be calculated by (GELMAN *et al.*, 2014)

$$\hat{n}_{eff} = \frac{mn}{1 + 2\sum_{t=1}^{T} \widehat{\rho}_t}, \quad \text{where} \quad \widehat{\rho}_t = 1 - \frac{V_t}{2\text{var}}, \quad V_t = \frac{1}{m(n-t)} \sum_{j=1}^{m} \sum_{i=t+1}^{n} (\theta_{i,j} - \theta_{i-t,j})^2. \tag{5.31}$$

5.5 Application to the stochastic reconstruction algorithm

In the previous sections, we described the statistical methods used in this work to derive estimators for the parameters of simulator-based models. As already mentioned, the stochastic reconstruction algorithm falls into this category. In this section, we intend to connect the described methods with our algorithm and define the main quantities such as the number of parameters, summary statistics and so forth.

Parameters

In Chapter 4, we defined the probability density functions used to model chemical attributes. We also introduced their functional forms and number of parameters. Here, we will take the discussion further. For distributions with the histogram functional (refer to Section 4.3), the parameters represent cumulative probabilities of the respective distribution. The cumulative probability is related to each parameter by

$$p_1 = \theta_j \tag{5.32}$$

$$p_{(1<i<n)} = \theta_i(1 - p_{i-1}) + p_{i-1} \tag{5.33}$$

$$p_n = 1 \tag{5.34}$$

where θ_j represents the first parameter of the distribution, p_1 represents the first possible outcome and p_n the last possible outcome. The number of parameters for these distributions is the number of possible outcomes minus one. It is clear that the parameters of the histograms are bounded to be between 0 and 1. In contrast, the distributions with known functional forms (chi-squared and exponential) have parameters controlling their overall shape. These parameters are bounded to be positive.

In Table 4.2, we showed a summary of the distributions' functional forms and the total number of parameters. As one can see, we defined a total of 14 parameters. However, based on the expected parameter identifiability we considered a few simplifications.

Distribution 9 (number of heterocycles) models the probability of the occurrence of heteroatoms. That is exactly what distributions 10, 11, and 12 models. The difference is in the functional forms of the heteroatoms. Distribution 9 models the cyclic form of heteroatoms and distributions 10, 11, and 12 models the aliphatic form of heteroatoms. However, experimental data reports only total amounts of heteroatoms. In that sense, we considered that distribution 9 and distributions 10, 11, and 12 share the same parameters. We have also considered that distribution 2 (length of a paraffinic chain) has the same parameter as distribution 7 (length of the side chain). In Table 5.1, we show the distributions and the parameters associated with each one. We have a total of 9 parameters to be estimated. Our framework has substantially fewer parameters than most works in the literature. In DE OLIVEIRA *et al.* (2013), the proposed model had a total of 24 parameters.

Table 5.1: Distributions and parameters labels and relationships

Chemical attribute	**Functional form**	**Available data?**	**Parameters**
Molecular type	Histogram	yes	θ_1, θ_2, θ_3
Lenght of a paraffinic chain	Chi-squared	yes	θ_4
Level of branching	Uniform	no	-
Total number of rings	Chi-squared	yes	θ_5
Ring configuration	Uniform	no	-
Methyl ring substitution	Uniform	no	-
Lenght of the side chain	Exponential	yes	θ_4
Number of benzenes	Exponential	yes	θ_6
Type of heterocycle	Histogram	yes	θ_7, $1-\theta_7$, θ_8
Aliphatic sulfur	Histogram	yes	θ_7
Another aliphatic heteroatom	Histogram	yes	$1-((1-\theta_7)(1-\theta_7)+\theta_7)$
Aliphatic nitrogen or oxygen	Histogram	yes	θ_8
Aliphatic oxygen function	Uniform	no	-
Number of cores	Exponential	yes	θ_9
Connection between cores	Uniform	no	-

Summary statistics

The stochastic reconstruction algorithm generates chemical attributes as simulated data. As discussed in Section 5.2, it is common to use summary statistics to represent data in approximate Bayesian computation algorithms. In our case, the use of summary statistics is necessary. The calculated properties from the hypothetical mixture is a form of summary statistics. Overall properties are the only form of observed data available. In that sense, we choose the summary statistics according to the available experimental data.

Discrepancy

Besides the definition of summary statistics, one of the main parts of the approximate Bayesian computation algorithms is the definition of the discrepancy function. This function measures the similarity between simulated and observed data. In this work, we chose the Mahalanobis discrepancy measure, defined by

$$d(y_\theta, y_0) = \left[(S(y_\theta) - S(y_0))^\intercal V_s^{-1} (S(y_\theta) - S(y_0))\right]^{1/2} \tag{5.35}$$

where $S()$ represents the summary statistics from simulated (and observed) data, and V_s represents a covariance matrix associated with the summary statistics.

Model reparametrization

Working with bounded parameters might cause convergence problems to the Hamiltonian Monte Carlo algorithm. All of our parameters are bounded. Part of them are bounded between 0 and 1, and part of them are bounded to be positive. In order to avoid convergence problems, we reparametrized our model. We used a new set of parameters φ, to be used in the estimation procedure. Those parameters are then rescaled within the model. For parameters bounded between 0 and 1 we used the logit function, and for parameters bounded to be positive we used the logarithm function, defined as follows

$$\theta_i = \frac{1}{1 + \exp(\varphi_i)}, \quad \text{for} \quad \theta_i = \{0, 1\}, \tag{5.36}$$

$$\theta_i = \exp(\varphi_i) \quad \text{for} \quad \theta_i > 0. \tag{5.37}$$

In this section, we defined the total number of parameters to be estimated by the statistical methods presented in Sections 5.3 and 5.4. We also defined the summary statistics and the discrepancy function. The reparametrization of the model is used to improve convergence. In that sense, the estimation procedure is done on the modified parameter set φ. This concludes the first part of the algorithm, estimation of the probability density functions parameters. In the next Sections, we will describe the methods applied to composition calculation by entropy maximization and molecular selection by clustering analysis.

5.6 Reconstruction by entropy maximization

In the previous sections, we described the methods used to estimate the parameters of the probability density distributions of the stochastic reconstruction algorithm. As showed in Section 4.5, in the stochastic reconstruction algorithm we considered that every built molecule has the same importance in the mixture. In other words, we considered an equimolar mixture. After estimating the parameters for the stochastic reconstruction algorithm, one can generate molecules and then calculate their composition based on available experimental data. In this work, we used the method developed by HUDEBINE and VERSTRAETE (2011), called reconstruction by entropy maximization.

Reconstruction by entropy maximization is based on the concept of information entropy, as proposed by SHANNON (1948). The entropy of a probability distribution is given by

$$E = -\sum_{i=1}^{N} p_i \log p_i, \tag{5.38}$$

where N is the number of points in the discrete probability distribution p. One can see the molecular composition x as a discrete probability distribution. In terms of molecular composition, the entropy Equation becomes

$$H = -\sum_{i=1}^{N} x_i \log x_i, \tag{5.39}$$

in this case, N represents the number of molecules and x_i each individual mole fraction.

The challenge of estimating the composition, or any probability distribution, is the lack of degrees of freedom. The number of points in the distribution N is usually much higher than the number of available experimental information. As a consequence, there are a number of different distributions that satisfies the observed data. Among the possible probability distributions, the least biased one is the one with maximum entropy (JAYNES; PRESSÉ *et al.*; SHANNON, 1957; 2013; 1948). Without any observed data, we expect the least biased distribution to be uniform. To prove that concept, we will apply the maximum entropy criteria directly to Equation 5.39. The only constraint is that the composition should sum to 1.

In order to maximize Equation 5.39, we need to add the constraint that all composition should sum to 1. The constrained entropy Equation becomes

$$H = -\sum_{i=1}^{N} x_i \log x_i + \tau(1 - \sum_{i=1}^{N} x_i), \tag{5.40}$$

where τ represents a Lagrangian multiplier associated with the imposed constraint. Taking the first and second derivatives in relation to x_i

$$\frac{\partial H}{\partial x_i} = -1 - \log x_i - \tau, \tag{5.41}$$

$$\frac{\partial^2 H}{\partial x_i^2} = \frac{-1}{x_i}. \tag{5.42}$$

Since x_i is always positive, by Equation 5.42, Equation 5.40 has an inflection point at its maximum. At the inflection point, we can solve Equation 5.41 for x_i,

$$x_i = \frac{1}{e^{1+\tau}}, \tag{5.43}$$

Equation 5.43 can be rewriteen as

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \frac{1}{e^{1+\tau}}, \tag{5.44}$$

solving for $e^{1+\tau}$ and substituing into Equation 5.43 yields

$$e^{1+\tau} = N, \tag{5.45}$$

$$x_i = \frac{1}{N}. \tag{5.46}$$

As expected, when no constraints are imposed, the distribution that maximizes the entropy is uniform. At this point, we need to discuss the type of constraints that we can use to deviate from the uniform distribution. In this work, we considered only exact linear constraints to the entropy Equation. These constraints have the following form

$$g_j = \sum_{i=1}^{N} x_i g_{ij}, \tag{5.47}$$

where g_j is an observed value for constraint j and g_{ij} is the contribution of molecule i to constraint j. For example, if g_j is the observed average molecular mass, g_{ij} is the molecular mass for each molecule i. By adding J constraints, corresponding to J observed data, Equation 5.40 becomes

$$H = -\sum_{i=1}^{N} x_i \log x_i + \tau(1 - \sum_{i=1}^{N} x_i) + \sum_{j=1}^{J} \lambda_j (g_j - \sum_{i=1}^{N} x_i g_{ij}), \tag{5.48}$$

where λ_j is a Lagrangian multiplier associated with constraint j, and J is the total number of constraints. By assuming that the constraints are linear, we are actually saying that g_{ij} is independent of x_i. The first derivative in relation to x_i is then

$$\frac{\partial H}{\partial x_i} = -1 - \log x_i - \tau - \sum_{j=1}^{J} \lambda_j g_{ij}, \tag{5.49}$$

in the inflection point, we can solve Equation 5.49 for x_i

$$0 = -1 - \log x_i - \tau - \sum_{j=1}^{J} \lambda_j g_{ij}, \tag{5.50}$$

$$e^{1+\tau} e^{(\log x_i)} = e^{(-\sum_{j=1}^{J} \lambda_j g_{ij})}, \tag{5.51}$$

$$x_i = \frac{e^{(-\sum_{j=1}^{J} \lambda_j g_{ij})}}{e^{1+\tau}}, \tag{5.52}$$

rearranging Equation 5.52, we can solve for $e^{1+\tau}$,

$$\sum_{i=1}^{N} x_i = \frac{1}{e^{1+\tau}} \sum_{i=1}^{N} e^{(-\sum_{j=1}^{J} \lambda_j g_{ij})}, \tag{5.53}$$

$$e^{1+\tau} = \sum_{i=1}^{N} e^{(-\sum_{j=1}^{J} \lambda_j g_{ij})}, \tag{5.54}$$

$$x_i = \frac{e^{(-\sum_{j=1}^{J} \lambda_j g_{ij})}}{Z}, \tag{5.55}$$

$$Z = \sum_{i=1}^{N} e^{(-\sum_{j=1}^{J} \lambda_j g_{ij})}, \tag{5.56}$$

substituing Equation 5.55 into Equation 5.48,

$$H = -\sum_{i=1}^{N} x_i \left[\log\left(e^{(-\sum_{j=1}^{J}\lambda_j g_{ij})}\right) - \log Z\right] + \sum_{j=1}^{J}\lambda_j(g_j - \sum_{i=1}^{N} x_i g_{ij}), \tag{5.57}$$

$$H = \log Z + \sum_{j=1}^{J}\lambda_j g_j. \tag{5.58}$$

From Equation 5.55, we see that the probability distribution, or in this case composition, that maximizes the entropy depends only on the Lagrangian multipliers λ_j for each constraint j. We can estimate λ_j by finding the extrema of Equation 5.58 in relation to λ_j.

In order to estimate molecular composition by the entropy maximization method we need to define the quantities g_{ij} and g_j. Clearly, the number of constraints j matches the number of observed data. Let g_{ij} be an element of a matrix $\mathbf{g}$, where each row represents a molecule and each column an observed experimental data (overall properties). When deriving Equations 5.55 and 5.58, we assumed that g_{ij} is independent of x_i. This is not always the case for the properties considered in this work.

To overcome this issue, we used the framework proposed in HUDEBINE and VERSTRAETE (2011). First, we consider g_j to be zero. In that sense, g_{ij} becomes a deviation between the observed value and the calculated value for constraint j. For example, for the average molecular mass we have that

$$g_{ij} = MW_{avg}^{obs} - \mathbf{MMW_i}, \tag{5.59}$$

where MW_{avg}^{obs} is the observed value of the average molecular mass and $\mathbf{MMW_i}$ is the molecular mass of molecule i (refer to Section 4.5). Combining Equations 5.47 and 5.59 leads to

$$0 = \sum_{i=1}^{N} x_i \left(MW_{avg}^{obs} - \mathbf{MMW_i}\right), \tag{5.60}$$

$$MW_{avg}^{obs} = \sum_{i=1}^{N} x_i \mathbf{MMW_i}, \tag{5.61}$$

which is exactly the type of constraint that we are looking for: the observed value matches the calculated value. In Equation 5.61, we can see that g_{ij} should be a

quantity that when multiplied by x_i and then $\sum_{i=1}^{N}$ results in the desired constraint. In that sense, we will derive g_{ij} for every available data considered in this work. The methodology for the calculation of the overall properties from the hypothetical mixture was already described in Section 4.5.

Elemental analysis: For an atom type k, this property can be calculated as

$$w_{k_{mix}} = \sum_{i=1}^{N} \left(\frac{x_i \mathbf{MA_{i,k}} \mathbf{VMW_{1,k}}}{\sum_{i=1}^{N} x_i \mathbf{MMW_i}} \right), \tag{5.62}$$

the constraint associated to g_{ij} is then calculated by

$$0 = \sum_{i=1}^{N} x_i \left(w_{k_{mix}}^{obs} - \frac{\mathbf{MA_{i,k}} \mathbf{VMW_{1,k}}}{\sum_{i=1}^{N} x_i \mathbf{MMW_i}} \right), \tag{5.63}$$

$$0 = \sum_{i=1}^{N} x_i \left(\frac{w_{k_{mix}}^{obs} \mathbf{MMW_i} - \mathbf{MA_{i,k}} \mathbf{VMW_{1,k}}}{\sum_{i=1}^{N} x_i \mathbf{MMW_i}} \right). \tag{5.64}$$

In Equation 5.64, we can see that the term g_{ij} is dependent of x_i. However, the limiting case where the calculated value matches with the observed value will occur if, and only if, the difference in the numerator reaches 0. In that sense, we can rewrite Equation 5.64 independent of x_i as

$$0 = \sum_{i=1}^{N} x_i \left(w_{k_{mix}}^{obs} \mathbf{MMW_i} - \mathbf{MA_{i,k}} \mathbf{VMW_{1,k}} \right), \tag{5.65}$$

$$g_{ij} = w_{k_{mix}}^{obs} \mathbf{MMW_i} - \mathbf{MA_{i,k}} \mathbf{VMW_{1,k}}. \tag{5.66}$$

Nuclear magnetic resonance: For a functional group of type k, its mole fraction in the mixture is given by

$$x_{k_{mix}} = \left(\frac{\sum_i x_i \mathbf{MF_{i,k}}}{\sum_i x_i \mathbf{MA_{i,k}}} \right), \tag{5.67}$$

the constraint and the associated g_{ij} are then calculated as

$$0 = \sum_{i=1}^{N} x_i \left(x_{k_{mix}}^{obs} \mathbf{MA_{i,k}} - \mathbf{MF_{i,k}} \right), \tag{5.68}$$

$$g_{ij} = x_{k_{mix}}^{obs} \mathbf{MA_{i,k}} - \mathbf{MF_{i,k}}. \tag{5.69}$$

SARA fractions: The total mass fraction of a group p in the mixture is the sum of the mass fractions of all molecules i that belongs to group p.

$$0 = \sum_{i=1}^{N} x_i \left(w_{p_{mix}}^{obs} - \frac{\mathbf{MMW_i}}{\sum_{i=1}^{N} x_i \mathbf{MMW_i}} \right), \tag{5.70}$$

$$0 = \sum_{i=1}^{N} x_i \left(\frac{\mathbf{MMW_i} \left(w_{p_{mix}}^{obs} - 1 \right)}{\sum_{i=1}^{N} x_i \mathbf{MMW_i}} \right), \tag{5.71}$$

$$0 = \sum_{i=1}^{N} x_i \left(\mathbf{MMW_i} \left(w_{p_{mix}}^{obs} - 1 \right) \right), \tag{5.72}$$

$$g_{ij} = \begin{cases} \mathbf{MMW_i} \left(w_{p_{mix}}^{obs} - 1 \right), & \text{if } i \text{ belongs to } p \\ \mathbf{MMW_i} \left(w_{p_{mix}}^{obs} \right), & \text{otherwise} \end{cases} \tag{5.73}$$

Distillation curve: consider an observed boiling point temperature $T_{b_k}^{obs}$, associated with a cumulative mass percent vaporization cw_k^{obs}. The calculated cumulative mass percent vaporization is the sum of the mass percentages of the molecules i that satisfies $T_{b,i} < T_{b_k}^{obs}$.

$$0 = \sum_{i=1}^{N} x_i \left(\frac{\mathbf{MMW_i}}{\sum_{i=1}^{N} x_i \mathbf{MMW_i}} - cw_k^{obs} \right), \tag{5.74}$$

$$0 = \sum_{i=1}^{N} x_i \left(\frac{\mathbf{MMW_i} \left(1 - cw_k^{obs} \right)}{\sum_{i=1}^{N} x_i \mathbf{MMW_i}} \right), \tag{5.75}$$

$$0 = \sum_{i=1}^{N} x_i \left(\mathbf{MMW_i} \left(1 - cw_k^{obs} \right) \right), \tag{5.76}$$

$$g_{ij} = \begin{cases} \mathbf{MMW_i} \left(1 - cw_k^{obs} \right), & \text{if } T_{b,i} < T_{b_k}^{obs} \\ -\mathbf{MMW_i} \left(cw_k^{obs} \right), & \text{otherwise} \end{cases} \tag{5.77}$$

Specific gravity: Given an observed specific gravity $S_{g_{mix}}^{obs}$, the associated constraint and g_{ij} is given by

$$0 = \sum_{i=1}^{N} x_i \left(\frac{\mathbf{MMW_i}}{\sum_{i=1}^{N} x_i \mathbf{MMW_i}} \frac{1}{d_i} - \frac{1}{S_{g_{mix}}^{obs}} \right), \tag{5.78}$$

$$0 = \sum_{i=1}^{N} x_i \left(\frac{\mathbf{MMW_i} \left(\frac{1}{d_i} - \frac{1}{S_{g_{mix}}^{obs}} \right)}{\sum_{i=1}^{N} x_i \mathbf{MMW_i}} \right), \tag{5.79}$$

$$0 = \sum_{i=1}^{N} x_i \left(\mathbf{MMW_i} \left(\frac{1}{d_i} - \frac{1}{S_{g_{mix}}^{obs}} \right) \right), \tag{5.80}$$

$$g_{ij} = \mathbf{MMW_i} \left(\frac{1}{d_i} - \frac{1}{S_{g_{mix}}^{obs}} \right). \tag{5.81}$$

In this section, we described the maximum entropy method applied to molecular composition calculation. We used the method proposed by HUDEBINE and VERSTRAETE (2011). One important result is the definition of the matrix $\mathbf{g}$, the constraint matrix used in Equation 5.58. As will be discussed in the next section, we used this matrix as the input for our clustering algorithm.

5.7 Partitioning around medoids

As described in Section 4.3, the stochastic reconstruction algorithm is performed by a Monte Carlo type sampling technique. For that reason, a large number of samples is necessary. In this work, we used a total of 5000 samples. Nothing guarantees that each sample forms a different molecule. The number of molecules can be limiting for certain types of applications. In that sense, we proposed a non-hierarchical clustering technique to select the best representative molecules from the overall molecular ensemble.

In this work, we used the algorithm proposed by PARK and JUN (2009). Consider N molecules each one having J variables. We intend to group them into $c\,(c < N)$ clusters, for a given c. We considered that the variables representing the molecules are the elements of the constraint matrix $\mathbf{g}$ defined in Section 5.6. In that sense, the jth variable of the molecule i is g_{ij}. A dissimilarity measure between molecules needs to be defined. In this work, we used the Euclidean distance to measure dissimilarity between molecule i and molecule i', as follows

$$d_{i,i'} = \left[\sum_{j=1}^{J} \left(g_{ij} - g_{i'j}\right)^2\right]^{1/2}. \tag{5.82}$$

With this method, we intend to find c medoids representing c clusters. As for any clustering algorithm, we expect that molecules within a cluster are similar but are dissimilar to molecules in other clusters (PARK and JUN, 2009). The algorithm used is described as follows (PARK and JUN, 2009):

- **Selection of initial medoids**
  - Calculate the distance between every pair of molecules.
  - Calculate the quantity $u_{i'}$ for molecule i' as follows

$$u_{i'} = \sum_{i=1}^{N} \frac{d_{i,i'}}{\sum_{i=1}^{N} d_{i,i'}}, \tag{5.83}$$

  - Select c molecules with the smallest values of u_j.
  - Obtain initial cluster by assigning each molecule to the nearest medoid.
  - Calculate the sum of distances from all objects to their medoids.
- **Update medoids**
  - Find a new medoid for each cluster by minimizing the total distance to other molecules in its cluster.
- **Assign molecules to medoids**
  - Assign each molecule to the nearest medoid.
  - Calculate the sum of distances from all objects to their medoids.
  - If the sum is equal to the previous one, stop the algorithm. Update medoids otherwise.

In this chapter, we described the statistical methods used in this work. First, we described the properties of simulator-based models and their consequences to parameter inference. We used the Bayesian optimization framework for approximate Bayesian computation. Second, we described the maximum entropy methodology for composition calculation. This methodology required the definition of a constraint matrix $\mathbf{g}$. At last, we used the constraint matrix $\mathbf{g}$ as an input for a non-hierarchical clustering technique. This technique is used to select c representative molecules called medoids. In the next chapter, we will discuss the applications of the proposed methodology.

Chapter 6

Results and discussions

In Chapter 4, we described the stochastic reconstruction algorithm developed in this work. Then, in Chapter 5, we discussed the statistical techniques used for parameter inference. The inference is divided into three steps. First, the probability density distributions parameters inference using Bayesian optimization is performed. Second, a clustering technique is applied to the built molecular ensemble to select c candidates representative of the mixture. Finally, the composition of the hypothetical mixture is calculated by entropy maximization.

The stochastic reconstruction model, the clustering technique and the reconstruction by entropy maximization method, were all implemented in this work. We used the C++ programming language with an object-oriented paradigm. For the Bayesian optimization framework, we used a Python package called *ELFI* (Engine for Likelihood-Free Inference) as described by JÄRVENPÄÄ *et al.* (2019).

In this chapter, we will describe the applications and results of the methods developed in this thesis. We applied our algorithm both to vacuum residues reported in the literature and for vacuum residues characterized at PETROBRAS research and development center.

6.1 Stochastic reconstruction

The first step of the algorithm is to estimate the parameters of the probability density functions controlling the molecular generation process. As described in Section 5.3, we used the Bayesian optimization framework for that purpose.

6.1.1 Model validation

One common practice of model validation in the approximate Bayesian computation is to fix the parameters and generate some data. Since the simulator intends to mimic the data generation process, we can use this simulated data as a pseudo

observation and then try to estimate the parameters previously fixed. In Table 6.1, we show the values for the fixed parameters, in the original and modified scale, and the summary statistics associated with the generated data. We choose this set of summary statistics based on the typical overall properties used in the literature. For the data generation process, we considered a total of 5000 samples (molecules). We performed 50 simulations with different seeds. Then, we calculated the mean and covariance matrix for the summary statistics. For sampling the variance surface, we used the Hamiltonian Monte Carlo algorithm with 50 iterations. For posterior sampling, we used the same algorithm with 4 chains for 1000 iterations each chain.

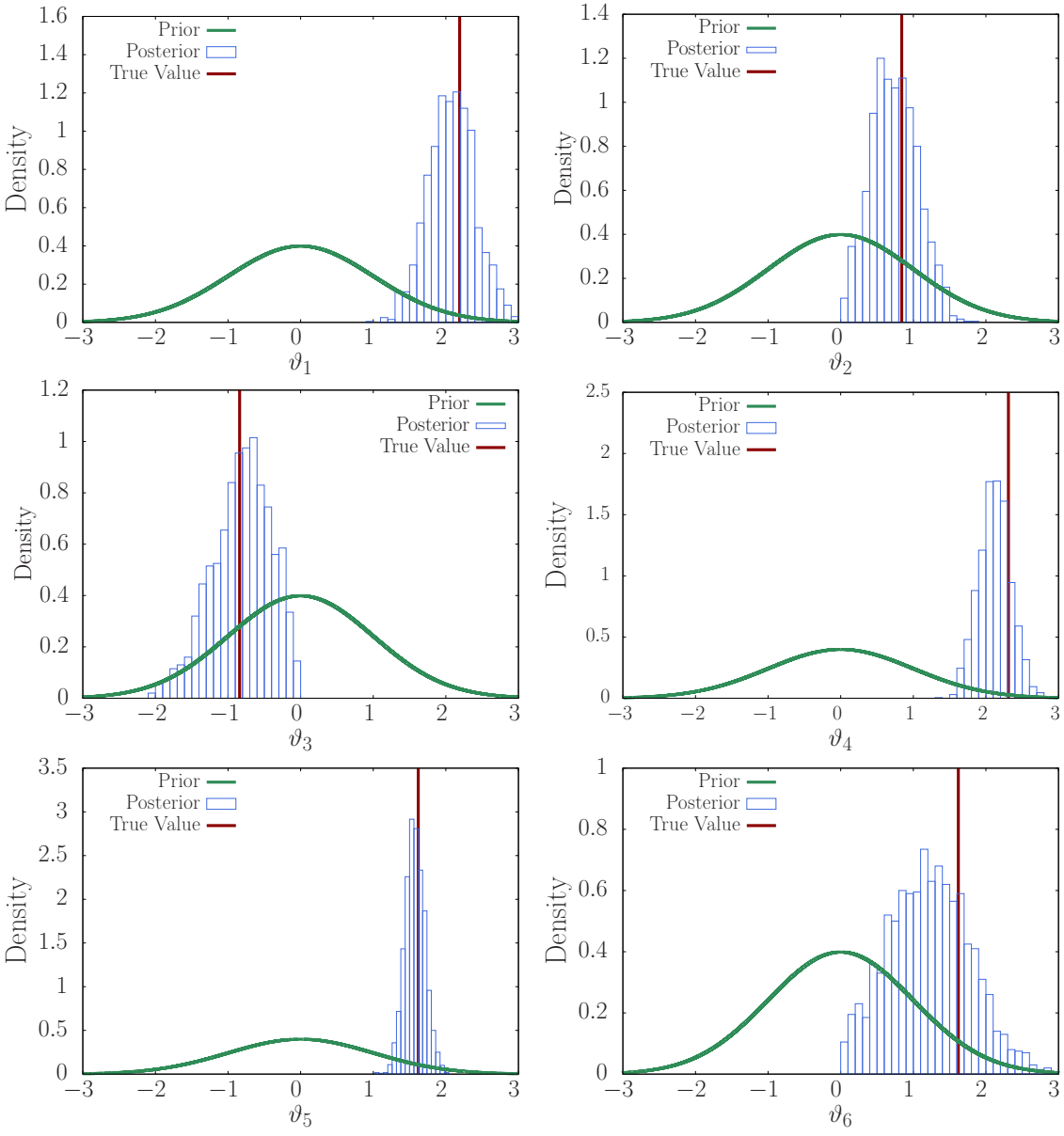
Table 6.1: Parameters and associated summary statistics for the model validation case.

Parameters (Original scale)		**Parameters (Rescaled)**		**Summary statistics (Pseudo-observed)**	
θ_1	0.1	ϑ_1	2.2	MW_{avg}^{obs} (g/mol)	640
θ_2	0.3	ϑ_2	0.85	**Elemental Analysis**	
θ_3	0.7	ϑ_3	-0.85	Carbon content (% w/w)	85.7
θ_4	10	ϑ_4	2.3	Hydrogen content (% w/w)	10.5
θ_5	5	ϑ_5	1.6	Sulfur content (% w/w)	1.6
θ_6	5	ϑ_6	1.6	Nitrogen content (% w/w)	0.61
θ_7	0.7	ϑ_7	-0.85	**SARA fractions**	
θ_8	0.3	ϑ_8	0.85	Saturates content (% w/w)	20.3
θ_9	2	ϑ_9	0.69	Aromatics content (% w/w)	28.1
				Resins content(% w/w)	35.2
				Nuclear magnetic resonance	
				Saturated carbon content (% m/m)	74.2
				Simulated distillation	
				Cummulative mass vaporization (%)	**Temperature ($^{\circ}C$)**
				10	413
				20	541
				30	651

In Figure 6.1, we show the posterior densities of the parameters for the validation case. We also show their prior densities and true values. All the parameters of the model are identifiable for the chosen set of summary statistics. In Table 6.2, we show the posterior mean and the 90% credible intervals of the parameters. We also show the convergence diagnosis quantities. The $100(1-\alpha)\%$ highest posterior density interval is the set of values that contains $100(1-\alpha)\%$ of the posterior probability and also has the characteristic that the density within the region is never lower than that outside (GELMAN *et al.*, 2014). All credible intervals contain the true values of the parameters.

Table 6.2: Posterior mean, credible intervals and convergence diagnosis of the parameters - Validation case

Parameters	**Posterior mean**	**Low 90% credible interval**	**High 90% credible interval**	$\hat{n}_{eff}$	$\hat{R}$
ϑ_1	2.08	1.59	2.66	3517.9	0.9999
ϑ_2	0.74	0.18	1.24	3827.7	0.9995
ϑ_3	-0.81	-1.49	-0.19	3524.2	1.0006
ϑ_4	2.12	1.76	2.48	4000	0.9999
ϑ_5	1.57	1.35	1.79	4000	0.9999
ϑ_6	1.22	0.27	2.11	1561.2	1.0016
ϑ_7	-0.82	-1.48	-0.11	2420.9	1.0005
ϑ_8	0.77	0.18	1.31	3088.5	1.0005
ϑ_9	0.76	0.19	1.3	3538.2	0.9996



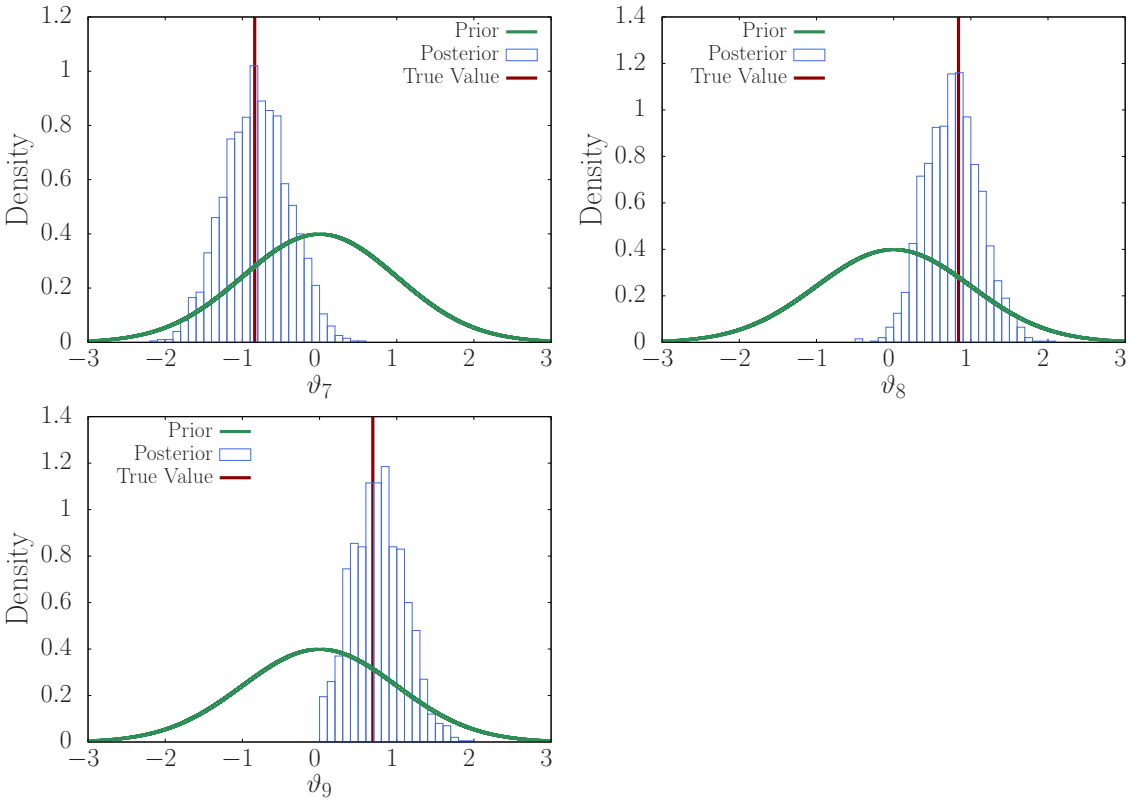
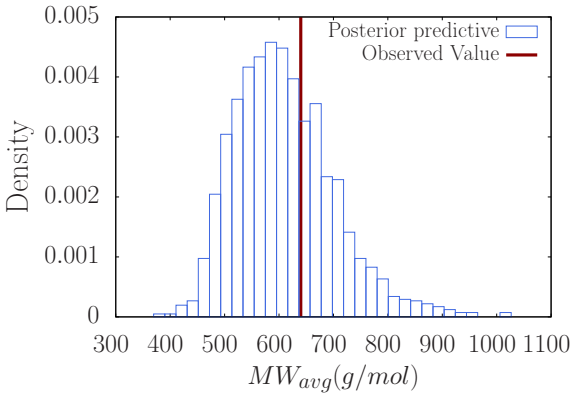


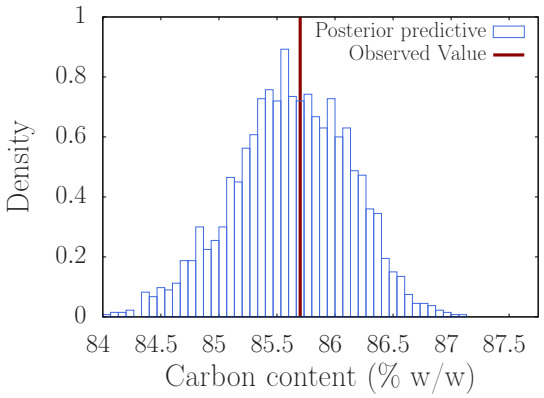
Figure 6.1: Prior and posterior densities, and true values of the parameters. Each graph represents a different parameter. The same prior was used to all parameters. True value of the parameters as shown in Table 6.1 - Validation case

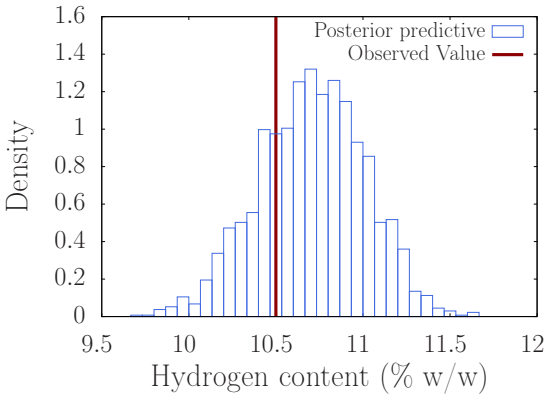
With the posterior density of the parameters, one can estimate the posterior predictive distribution of a new observation $\tilde{y}$ conditioned on the observed data y_0

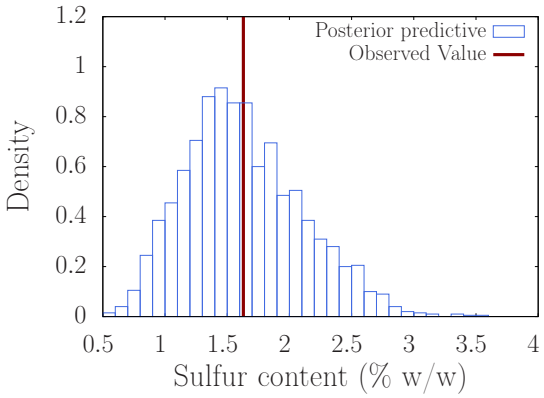
$$p(\tilde{y}|y_0) = \int p(\tilde{y}|\theta)p(\theta|y_0)d\theta. \tag{6.1}$$

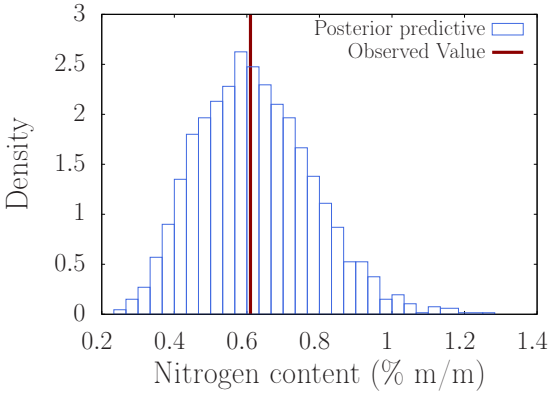
In practice, we take samples from the posterior distributions of parameters and run the simulator. The outputs of the simulator are samples of the posterior predictive distribution for any quantity of interest. For the validation case, we used the posterior predictive distribution to check if the model can replicate the observed summary statistics considered. In Figure 6.2, we show the posterior predictive distribution and observed value for the summary statistics considered. In Table 6.3, we show the posterior predictive mean and credible intervals. The model can replicate the observed summary statistics. All credible intervals contain the observed value and the posterior mean is a good estimator for the observed summary statistics. In that sense, we considered the model validated.

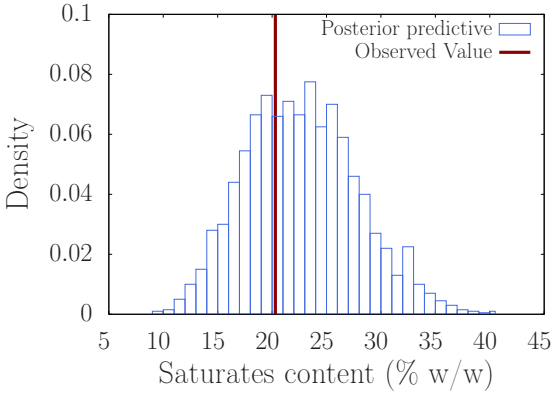

Posterior predictive
Observed Value
Density
0.005
0.004
0.003
0.002
0.001
0
300 400 500 600 700 800 900 1000 1100
$MW_{avg}(g/mol)$

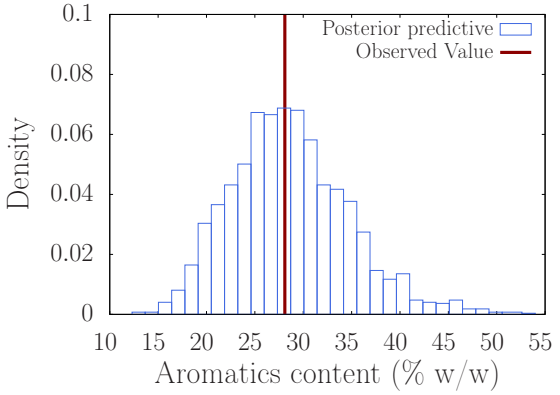

Posterior predictive
Observed Value
Density
1
0.8
0.6
0.4
0.2
0
84 84.5 85 85.5 86 86.5 87 87.5
Carbon content (% w/w)

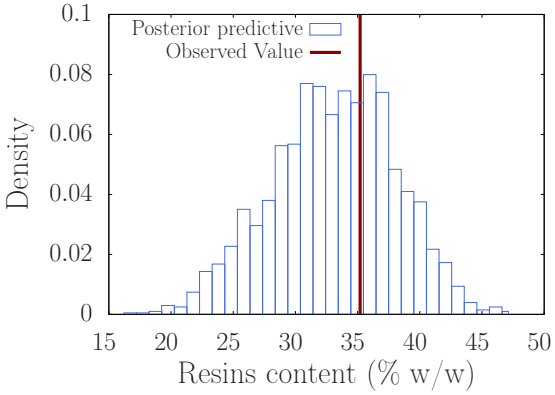

Posterior predictive
Observed Value
Density
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
9.5 10 10.5 11 11.5 12
Hydrogen content (% w/w)


Posterior predictive
Observed Value
Density
1.2
1
0.8
0.6
0.4
0.2
0
0.5 1 1.5 2 2.5 3 3.5 4
Sulfur content (% w/w)


Posterior predictive
Observed Value
Density
3
2.5
2
1.5
1
0.5
0
0.2 0.4 0.6 0.8 1 1.2 1.4
Nitrogen content (% m/m)


Posterior predictive
Observed Value
Density
0.1
0.08
0.06
0.04
0.02
0
5 10 15 20 25 30 35 40 45
Saturates content (% w/w)


Posterior predictive
Observed Value
Density
0.1
0.08
0.06
0.04
0.02
0
10 15 20 25 30 35 40 45 50 55
Aromatics content (% w/w)


Posterior predictive
Observed Value
Density
0.1
0.08
0.06
0.04
0.02
0
15 20 25 30 35 40 45 50
Resins content (% w/w)

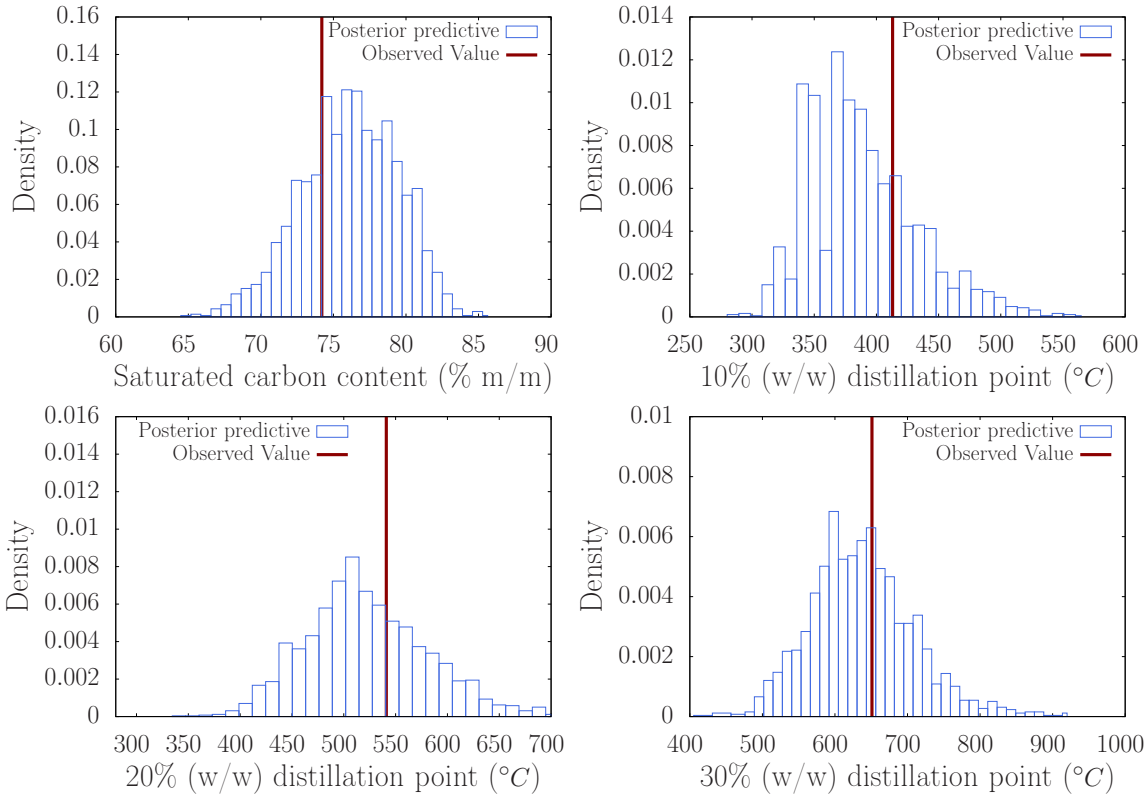


Figure 6.2: Posterior predictive distribution for the summary statistics (overall properties) used in the discrepancy function. Each graph represents a different propertie. The observed value is included for comparison - Validation case

Table 6.3: Posterior predictive distribution mean and credible intervals - Validation case.

Summary statistics	**Observed**	**Posterior mean**	**Low 90% credible interval**	**High 90% credible interval**
MW_{avg} (g/mol)	640	610	464	742
Elemental Analysis				
Carbon content (% w/w)	85.7	85.6	84.8	86.4
Hydrogen content (% w/w)	10.5	10.7	10.2	11.2
Sulfur content (% w/w)	1.6	1.6	0.8	2.4
Nitrogen content (% w/w)	0.61	0.62	0.36	0.86
SARA fractions				
Saturates content (% w/w)	20.3	22.7	13.8	30.5
Aromatics content (% w/w)	28.1	28.6	18.6	37.9
Resins content(% w/w)	35.2	33.1	25	41.1
Nuclear magnetic resonance				
Saturated carbon content (% m/m)	74.2	76.1	70.7	81.3
Simulated distillation Temperature (°C) at cumulative mass vaporization (%)				
10	413	389	326	457
20	541	522	429	620
30	651	638	507	733

6.1.2 Application to vacuum residues from different origins

In this work, we applied our model to four different vacuum residues from different origins. Two from DE OLIVEIRA *et al.* (2013) and two characterized at PETROBRAS research and development center. In Table 6.4, we show the overall properties (summary statistics) for the considered vacuum residues. In Table 6.5, we show additional properties for the two vacuum residues from PETROBRAS research and development center, which are results from a hydrogen nuclear magnetic resonance and some additional carbon types from carbon nuclear magnetic resonance. These properties were not included in the parameter estimation step. The comparison between the calculated and experimental values for the additional properties is another form of model validation.

Table 6.4: Summary statistics for two different vacuum residues from DE OLIVEIRA *et al.* (2013) and two characterized at PETROBRAS research and development center

Summary statistics	**Ural**	**Maya**	**Vacuum Residue A**	**Vacuum Residue B**
MW_{avg} (g/mol)	727	764	718	751
Elemental Analysis				
Carbon content (% w/w)	85.5	85.2	86.7	86.5
Hydrogen content (% w/w)	10.6	10.1	11.4	11.0
Sulfur content (% w/w)	2.7	3.5	0.6	0.7
Nitrogen content (% w/w)	0.58	0.58	1.0	0.9
SARA fractions				
Saturates content (% w/w)	11.7	12.9	19.0	12.0
Aromatics content (% w/w)	46.1	38.7	40.0	45.0
Resins content(% w/w)	37.6	34.2	34.0	33.0
Nuclear magnetic resonance				
Saturated carbon content (% m/m)	72.8	69.5	77.6	75.1
Simulated distillation Temperature ($^{\circ}C$) at cumulative mass vaporization (%)				
10	520	520	525	526
20	550	558	555	561
30	574	585	576	585

Table 6.5: Additional properties for the two vacuum residues characterized in PETROBRAS research and development center used for validation purposes.

Summary statistics	**Vacuum Residue A**	**Vacuum Residue B**
Nuclear magnetic resonance		
Aromatic hydrogen content (% m/m)	4.8	5.6
α - hydrogen content (% m/m)	8.5	9.8
β - hydrogen content (% m/m)	64.9	63.9
γ - hydrogen content (% m/m)	20.6	18.5
CH_3/paraffinic CH_2	0.3	0.32
Branched CH_3/paraffinic CH_2	0.2	0.22

In Tables 6.6 to 6.9, we show the estimated parameters for the vacuum residues considered in this work. The posterior distribution is represented in terms of the posterior mean and the 90% credible intervals. We also show the quantities for convergence diagnosis. Overall, the main differences are in the parameters controlling the molecular types and heteroatoms quantities. The parameters controlling the number of rings and length of the paraffinic chain are relatively close between the vacuum residues. Vacuum residues A and B have a higher paraffinics and naphthenics content compared to Ural and Maya. This is consistent with the experimental data, since vacuum residues A and B have higher hydrogen and saturated carbon content.

Table 6.6: Posterior mean, credible intervals and convergence diagnosis of the parameters - Ural

Parameters	**Posterior mean**	**Low 90% credible interval**	**High 90% credible interval**	$\hat{n}_{eff}$	$\hat{R}$
ϑ_1	1.55	1.1	2.04	3176.2	1.0005
ϑ_2	3.35	2.75	3.94	2140.2	0.9998
ϑ_3	-3.14	-3.87	-2.38	1663.8	1.0007
ϑ_4	4.25	3.87	4.61	3234.4	1.0002
ϑ_5	2.04	1.88	2.2	3247.5	1.0001
ϑ_6	2.62	2.07	3.34	2222.9	1.0005
ϑ_7	0.53	-0.04	1.02	2207.3	1.0014
ϑ_8	-0.09	-0.72	0.49	2459.7	1.0001
ϑ_9	-0.33	-1.28	0.73	1236.5	1.003

Table 6.7: Posterior mean, credible intervals and convergence diagnosis of the parameters - Maya

Parameters	**Posterior mean**	**Low 90% credible interval**	**High 90% credible interval**	$\hat{n}_{eff}$	$\hat{R}$
ϑ_1	2.11	1.58	2.66	2372.3	1.0005
ϑ_2	1.88	1.44	2.3	3383.2	1.0002
ϑ_3	-2.55	-3.19	-1.93	1857.8	1.0007
ϑ_4	3.58	3.27	3.9	3714.3	0.9997
ϑ_5	2.00	1.85	2.16	3084.9	1.0005
ϑ_6	3.31	2.56	4.04	1361.9	0.9994
ϑ_7	0.99	0.51	1.42	2805.3	1.0006
ϑ_8	0.04	-0.39	0.46	2969.2	1.0007
ϑ_9	0.08	-1.01	1.36	775.1	1.0064

Table 6.8: Posterior mean, credible intervals and convergence diagnosis of the parameters - Vacuum residue A

Parameters	**Posterior mean**	**Low 90% credible interval**	**High 90% credible interval**	$\hat{n}_{eff}$	$\hat{R}$
ϑ_1	0.83	0.50	1.14	4000	0.9994
ϑ_2	1.85	1.41	2.25	3601.6	0.9994
ϑ_3	-1.55	-2.0	-1.01	3174.9	1.0002
ϑ_4	3.69	3.37	4.01	3401.5	1.0001
ϑ_5	1.86	1.67	2.04	3171.9	1.0000
ϑ_6	2.61	2.01	3.32	1674.2	1.0002
ϑ_7	-1.41	-1.89	-0.83	2909.1	1.0000
ϑ_8	-1.11	-1.79	-0.50	1663	1.0006
ϑ_9	-0.41	-1.05	0.15	1818.6	1.0001

Table 6.9: Posterior mean, credible intervals and convergence diagnosis of the parameters -Vacuum residue B

Parameters	**Posterior mean**	**Low 90% credible interval**	**High 90% credible interval**	$\hat{n}_{eff}$	$\hat{R}$
ϑ_1	1.14	0.67	1.59	3524.5	1.0001
ϑ_2	2.66	2.1	3.23	3854.2	1.0004
ϑ_3	-1.81	-2.52	-1.11	3042.9	0.9996
ϑ_4	3.41	2.96	3.94	2586.9	0.9998
ϑ_5	1.91	1.7	2.1	3985.5	0.9993
ϑ_6	2.06	1.21	2.89	1785.5	1.0003
ϑ_7	-1.71	-2.49	-1.00	2698.4	1.0005
ϑ_8	-0.03	-0.63	0.56	1236.7	0.9992
ϑ_9	0.21	-0.76	1.22	1975.6	1.0007

With the posterior distribution of parameters, we calculated the posterior predictive distribution for the overall properties for each vacuum residue. We used all posterior samples of the parameters to generate the posterior predictive distribution. We divided the properties between the constrained (used in the discrepancy function for parameter estimation) and unconstrained. The use of unconstrained properties is another form of model validation.

In Tables 6.10 to 6.13, we show the posterior mean and 90% credible intervals for the summary statistics (properties) used in the parameter estimation step. The model was able to replicate most of the observed summary statistics (properties). However, for all vacuum residues, the distillation curve is not well replicated. This can be due to the ideal mixture consideration and/or the group contribution method for boiling point calculation. Nevertheless, the entropy maximization step corrects this discrepancy, as we will show in the next section.

In Tables 6.14 and 6.15, we show the posterior mean and 90% credible intervals for the unconstrained properties for vacuum residues A and B. As already mentioned, these properties were not used in the discrepancy function for the parameter estimation step. We can see a good agreement between prediction and observed values. This shows the model's ability to predict different properties from the generated molecular ensemble.

In Figures 6.3 to 6.8, we show a graphical representation of the posterior predictive distribution for the vacuum residues studied in this work. Besides the replicated summary statistics, we also show the additional predicted properties for vacuum residues A and B. Similarly to what is shown in Tables 6.10 to 6.15, the model can replicate the observed data, except for the distillation curve.

Overall, the model can represent the different vacuum residues studied in this work. Both constrained and unconstrained observed properties are contained within the posterior predictive distribution credible intervals. Besides that, the posterior predictive mean is a good estimator for the observed properties.

The Bayesian optimization framework is very effective in estimating the posterior distribution of the parameters. The use of the Bayesian statistics paradigm turns uncertainty propagation a natural process. Besides that, the use of prior distributions tends to avoid superparametrization issues.

Table 6.10: Posterior predictive distribution mean and credible intervals - Ural.

Summary statistics	**Observed**	**Posterior mean**	**Low 90% credible interval**	**High 90% credible interval**
MW_{avg} (g/mol)	727	752	670	832
Elemental Analysis				
Carbon content (% w/w)	85.5	85	84,4	85,7
Hydrogen content (% w/w)	10.6	10,6	10,2	10,9
Sulfur content (% w/w)	2.7	2,78	2,18	3,36
Nitrogen content (% w/w)	0.58	0,81	0,55	1,04
SARA fractions				
Saturates content (% w/w)	11.7	12,1	8,3	15,8
Aromatics content (% w/w)	46.1	45,6	38,3	53,2
Resins content(% w/w)	37.6	36,8	30,2	43,8
Nuclear magnetic resonance				
Saturated carbon content (% m/m)	72.8	73.6	70.7	76.6
Simulated distillation Temperature ($^{\circ}C$) at cumulative mass vaporization (%)				
10	520	560	460	656
20	550	683	629	729
30	574	744	696	795

Table 6.11: Posterior predictive distribution mean and credible intervals - Maya.

Summary statistics	**Observed**	**Posterior mean**	**Low 90% credible interval**	**High 90% credible interval**
MW_{avg} (g/mol)	764	742	634	832
Elemental Analysis				
Carbon content (% w/w)	85.2	84,7	84,1	85,3
Hydrogen content (% w/w)	10.1	10,1	9,8	10,4
Sulfur content (% w/w)	3.5	3,47	2,89	3,95
Nitrogen content (% w/w)	0.58	0,83	0,66	1,03
SARA fractions				
Saturates content (% w/w)	12.9	16,1	11,1	20,6
Aromatics content (% w/w)	38.7	35,8	28,9	41,8
Resins content(% w/w)	34.2	36,4	30,2	42,3
Nuclear magnetic resonance				
Saturated carbon content (% m/m)	69.5	70,7	67,9	73,3
Simulated distillation Temperature ($^{\circ}C$) at cumulative mass vaporization (%)				
10	520	566	506	637
20	528	667	610	719
30	585	731	682	789

Table 6.12: Posterior predictive distribution mean and credible intervals - Vacuum residue A.

Summary statistics	**Observed**	**Posterior mean**	**Low 90% credible interval**	**High 90% credible interval**
MW_{avg} (g/mol)	718	656	578	735
Elemental Analysis				
Carbon content (% w/w)	86.7	86.2	85.7	86.6
Hydrogen content (% w/w)	11.4	11.5	11.2	11.9
Sulfur content (% w/w)	0.6	0.83	0.49	1.19
Nitrogen content (% w/w)	1.0	1.07	0.86	1.28
SARA fractions				
Saturates content (% w/w)	19.0	21.4	16.3	25.8
Aromatics content (% w/w)	40.0	44.4	35.6	53.1
Resins content(% w/w)	34.0	31.9	24.1	40.1
Nuclear magnetic resonance				
Saturated carbon content (% m/m)	77.6	79.7	77	82.7
Simulated distillation Temperature ($^{\circ}C$) at cumulative mass vaporization (%)				
10	525	369	329	403
20	555	572	500	652
30	576	669	604	730

Table 6.13: Posterior predictive distribution mean and credible intervals - Vacuum residue B.

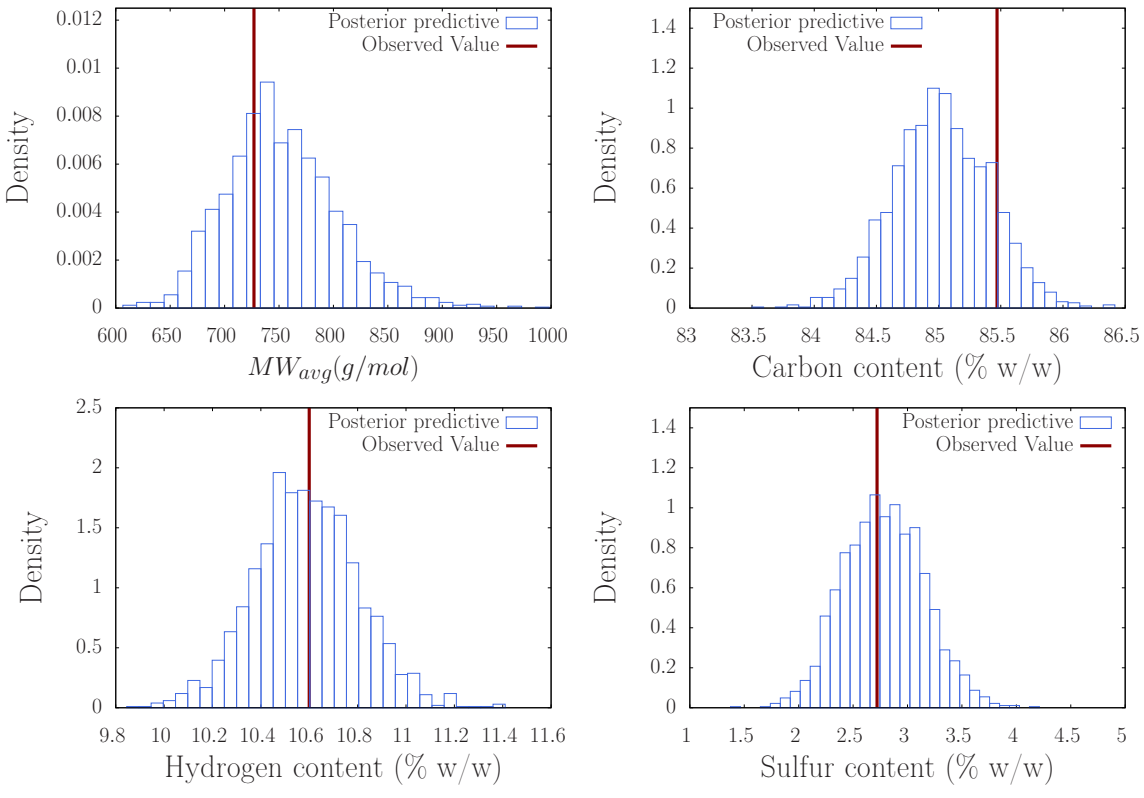
Summary statistics	**Observed**	**Posterior mean**	**Low 90% credible interval**	**High 90% credible interval**
MW_{avg} (g/mol)	751	699	521	834
Elemental Analysis				
Carbon content (% w/w)	86.5	86,4	85,8	87
Hydrogen content (% w/w)	11.0	11	10,6	11,6
Sulfur content (% w/w)	0.7	0,77	0,31	1,22
Nitrogen content (% w/w)	0.9	0,86	0,6	1,12
SARA fractions				
Saturates content (% w/w)	12.0	16,5	10,8	21,5
Aromatics content (% w/w)	45.0	42,5	32	54,2
Resins content(% w/w)	33.0	33	24,2	41,3
Nuclear magnetic resonance				
Saturated carbon content (% m/m)	75.1	75,9	71,6	80,3
Simulated distillation Temperature ($^{\circ}C$) at cumulative mass vaporization (%)				
10	526	425	330	525
20	561	599	508	691
30	585	690	603	774

Table 6.14: Posterior predictive distribution mean and credible intervals for the unconstrained properties - Vacuum residue A.

Summary statistics	Observed	Posterior mean	Low 90% credible interval	High 90% credible interval
Nuclear magnetic resonance				
Aromatic hydrogen content (% m/m)	4.8	5.7	4.6	6.9
α - hydrogen content (% m/m)	8.5	7.6	6.4	8.6
β - hydrogen content (% m/m)	64.9	65.3	63.3	67.7
γ - hydrogen content (% m/m)	20.6	21.4	19.5	23.2
CH_3/paraffinic CH_2	0.30	0.35	0.32	0.38
Branched CH_3/paraffinic CH_2	0.20	0.21	0.19	0.23

Table 6.15: Posterior predictive distribution mean and credible intervals for the unconstrained properties - Vacuum residue B.

Summary statistics	Observed	Posterior mean	Low 90% credible interval	High 90% credible interval
Nuclear magnetic resonance				
Aromatic hydrogen content (% m/m)	5.6	6.7	5.0	8.6
α - hydrogen content (% m/m)	9.8	8.6	7.0	10.5
β - hydrogen content (% m/m)	63.9	64.3	60.1	68.6
γ - hydrogen content (% m/m)	18.5	20.3	18.1	23.2
CH_3/paraffinic CH_2	0.32	0.36	0.30	0.41
Branched CH_3/paraffinic CH_2	0.22	0.21	0.18	0.24



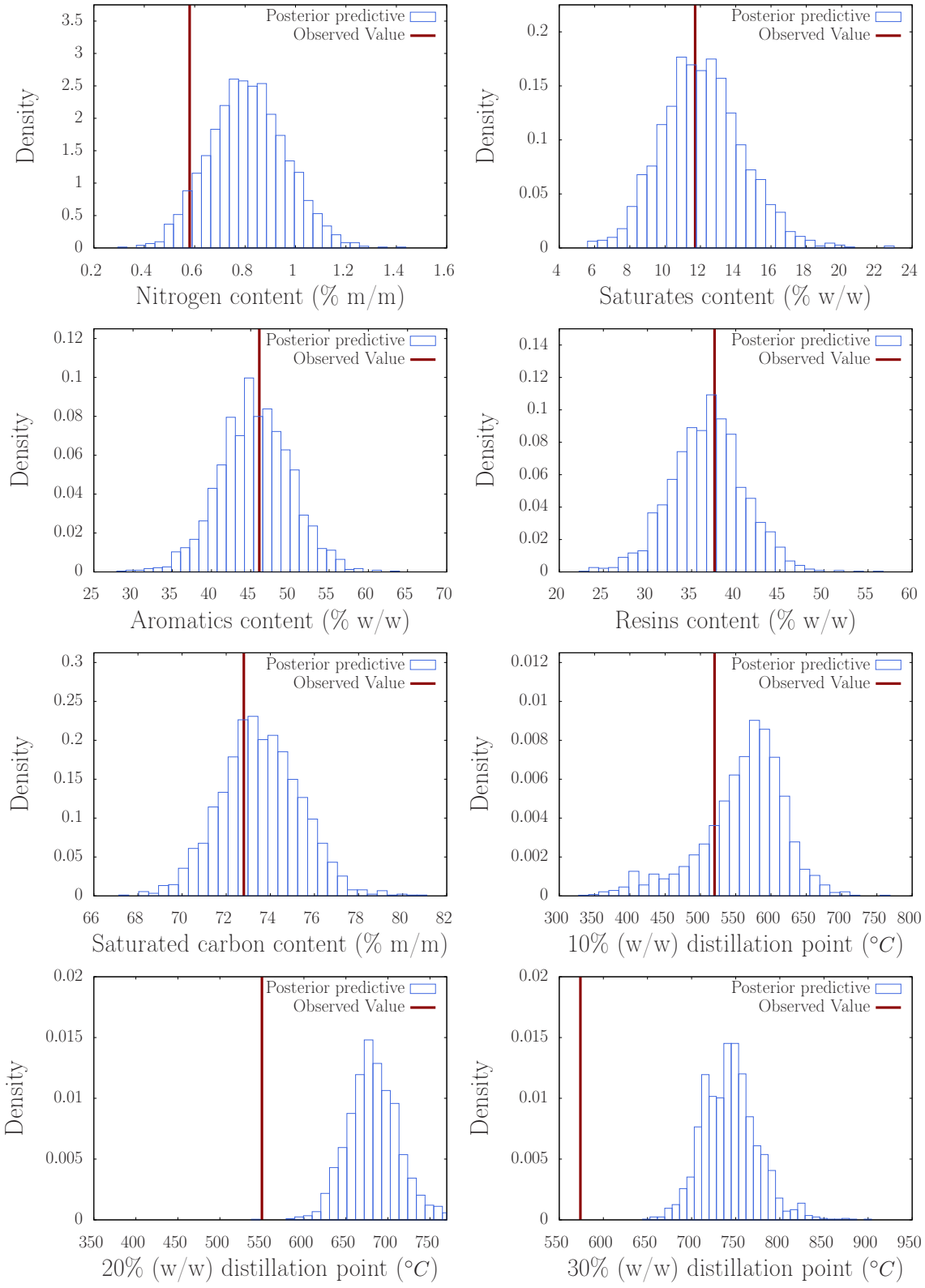
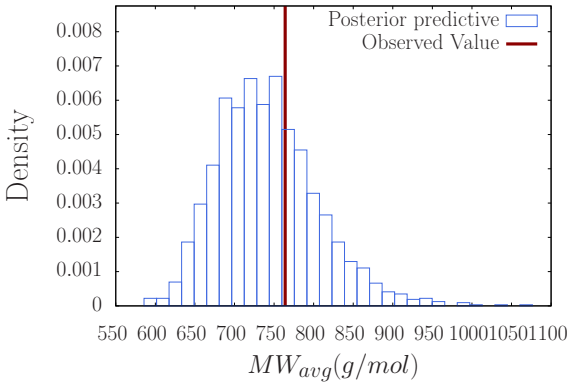
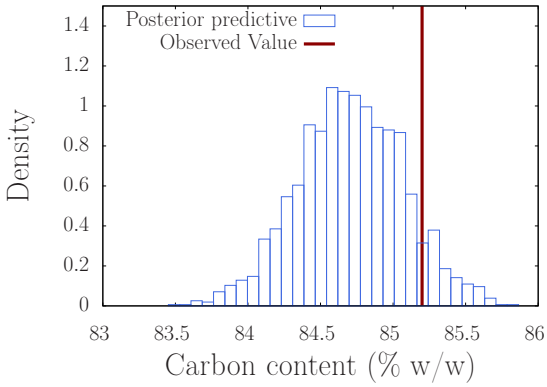
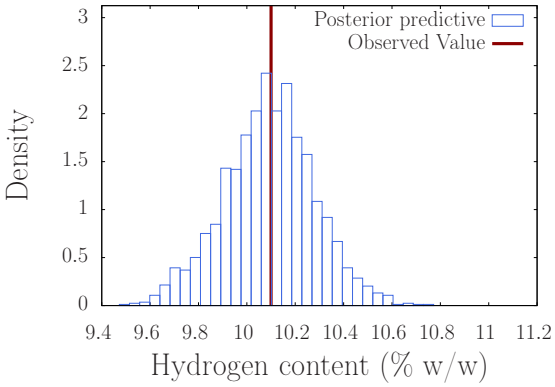
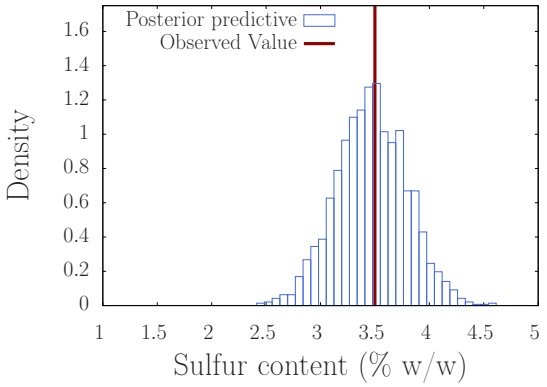


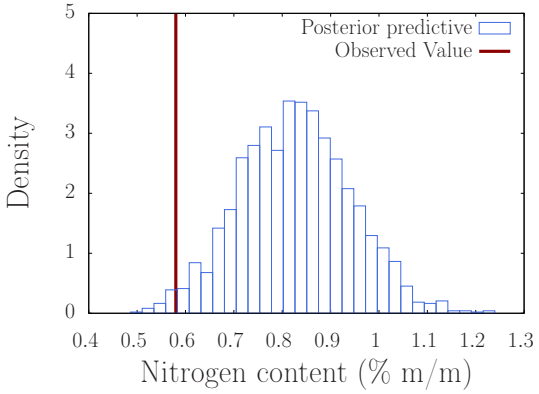
Figure 6.3: Posterior predictive distribution for the summary statistics (overall properties) used in the discrepancy function. Each graph represents a different property. The observed value is included for comparison. The model can replicate most of the observed data. An exception to the distillation curve - Vacuum residue Ural.

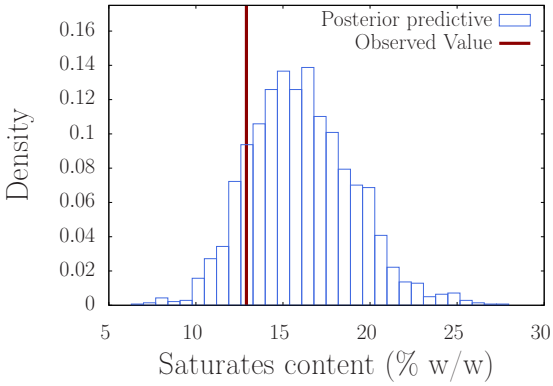

Posterior predictive
Observed Value
Density
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
550 600 650 700 750 800 850 900 950 1000 1050 1100
$MW_{avg}(g/mol)$

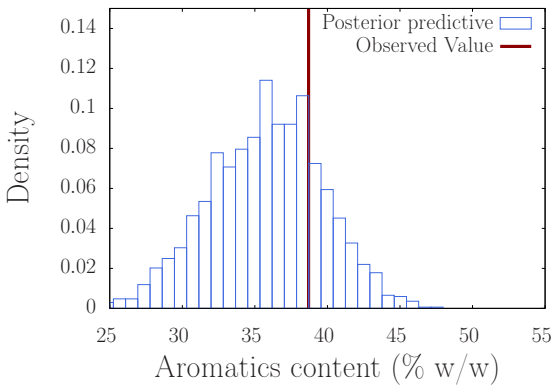

Posterior predictive
Observed Value
Density
1.4
1.2
1
0.8
0.6
0.4
0.2
0
83 83.5 84 84.5 85 85.5 86
Carbon content (% w/w)

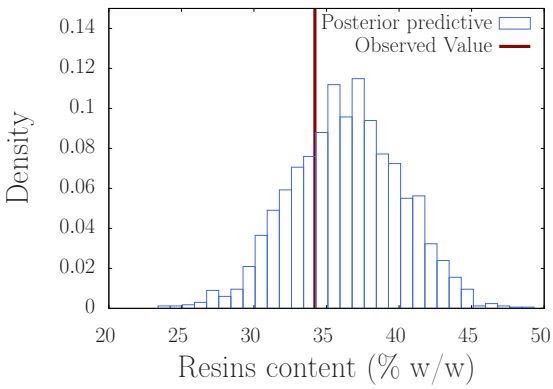

Posterior predictive
Observed Value
Density
3
2.5
2
1.5
1
0.5
0
9.4 9.6 9.8 10 10.2 10.4 10.6 10.8 11 11.2
Hydrogen content (% w/w)


Posterior predictive
Observed Value
Density
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1 1.5 2 2.5 3 3.5 4 4.5 5
Sulfur content (% w/w)


Posterior predictive
Observed Value
Density
5
4
3
2
1
0
0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3
Nitrogen content (% m/m)


Posterior predictive
Observed Value
Density
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
5 10 15 20 25 30
Saturates content (% w/w)


Posterior predictive
Observed Value
Density
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
25 30 35 40 45 50 55
Aromatics content (% w/w)


Posterior predictive
Observed Value
Density
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
20 25 30 35 40 45 50
Resins content (% w/w)

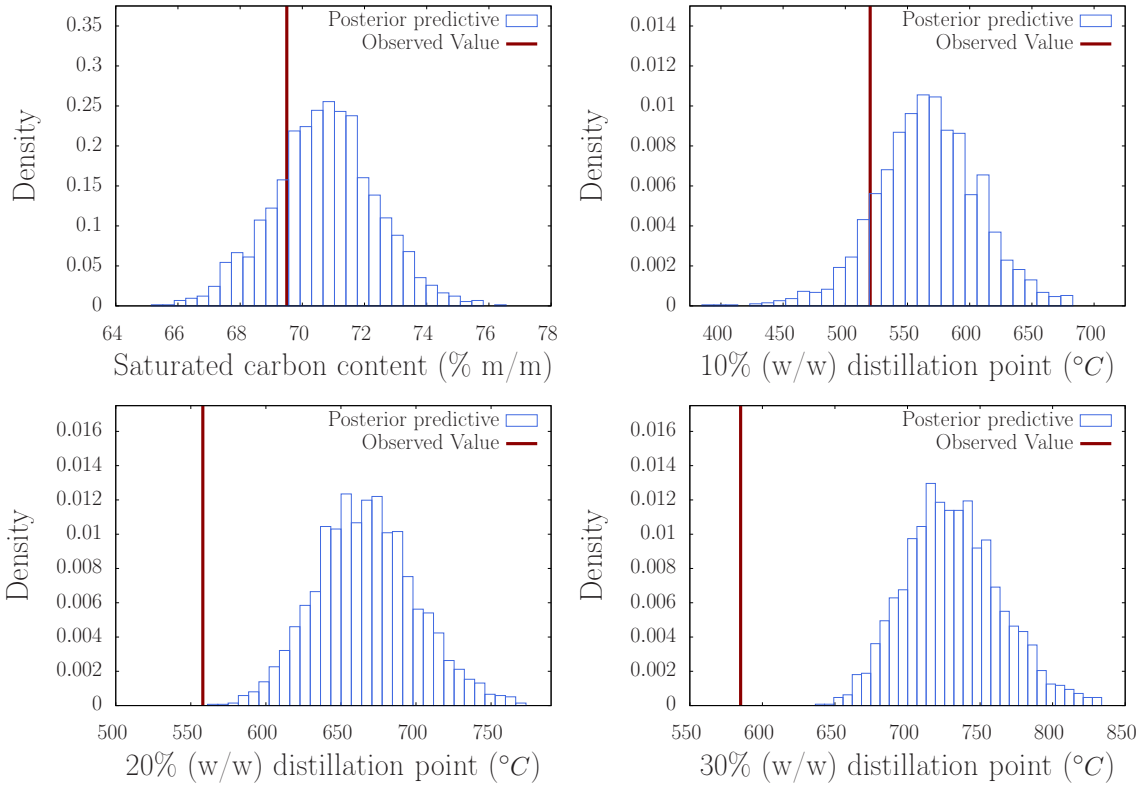
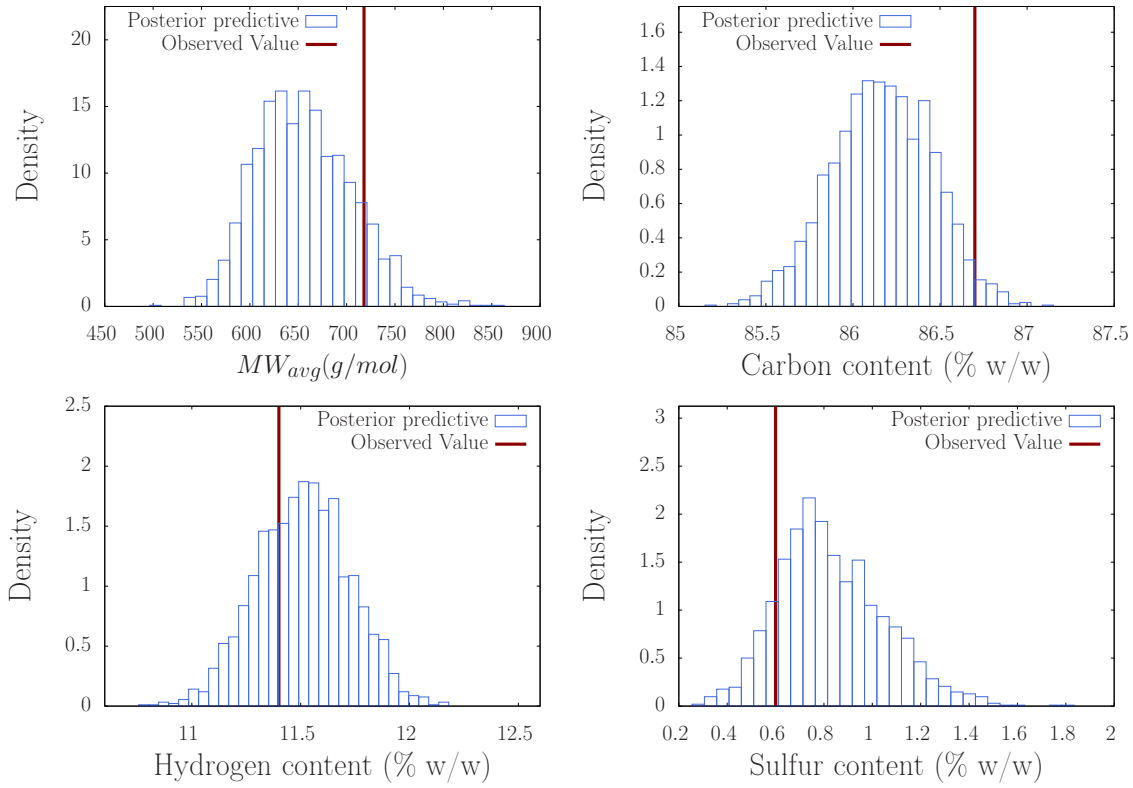


Figure 6.4: Posterior predictive distribution for the summary statistics (overall properties) used in the discrepancy function. Each graph represents a different property. The observed value is included for comparison. The model can replicate most of the observed data. An exception to the distillation curve - Vacuum residue Maya.



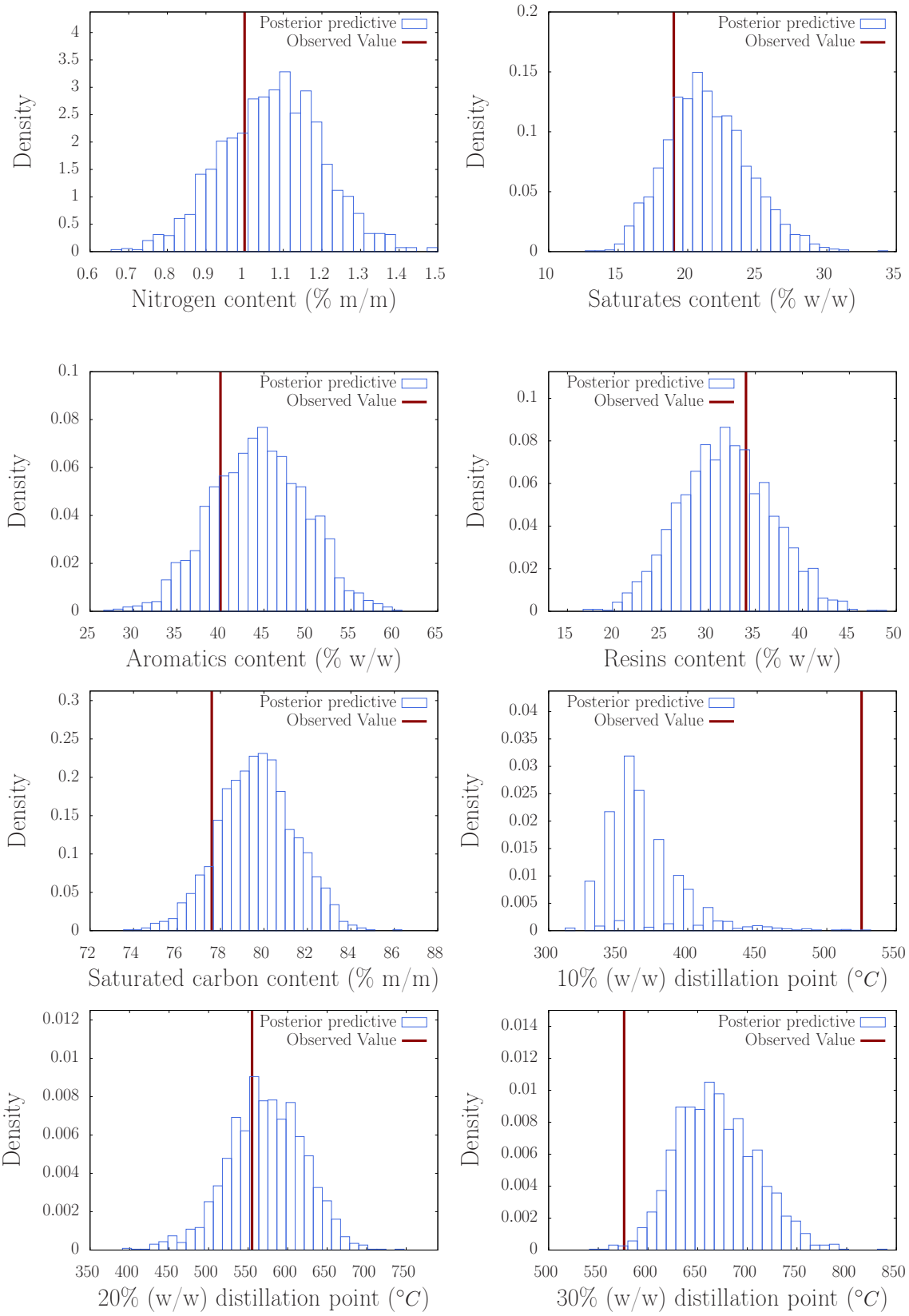


Figure 6.5: Posterior predictive distribution for the summary statistics (overall properties) used in the discrepancy function. Each graph represents a different property. The observed value is included for comparison. The model can replicate most of the observed data. An exception to the distillation curve - Vacuum residue A.

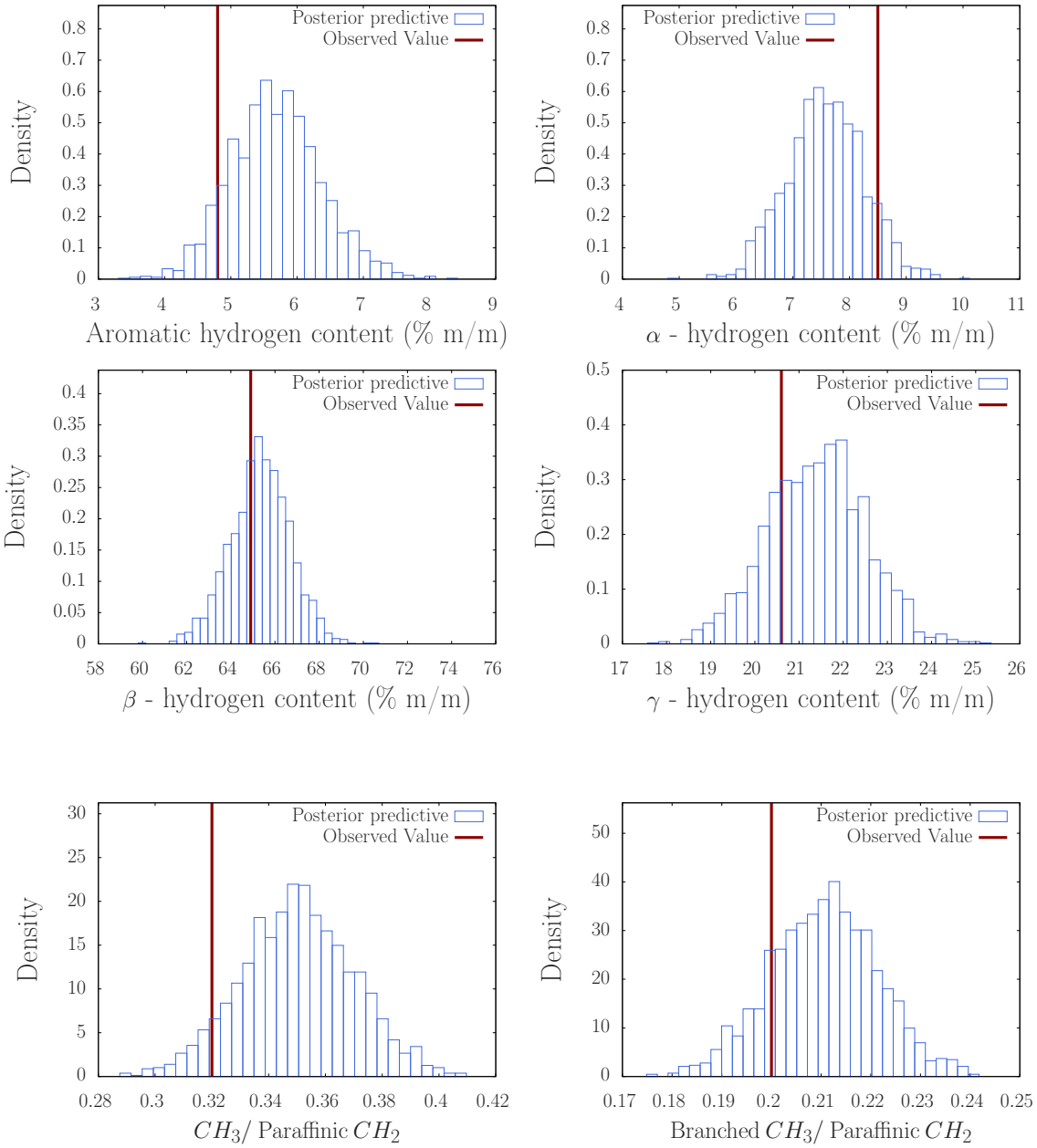
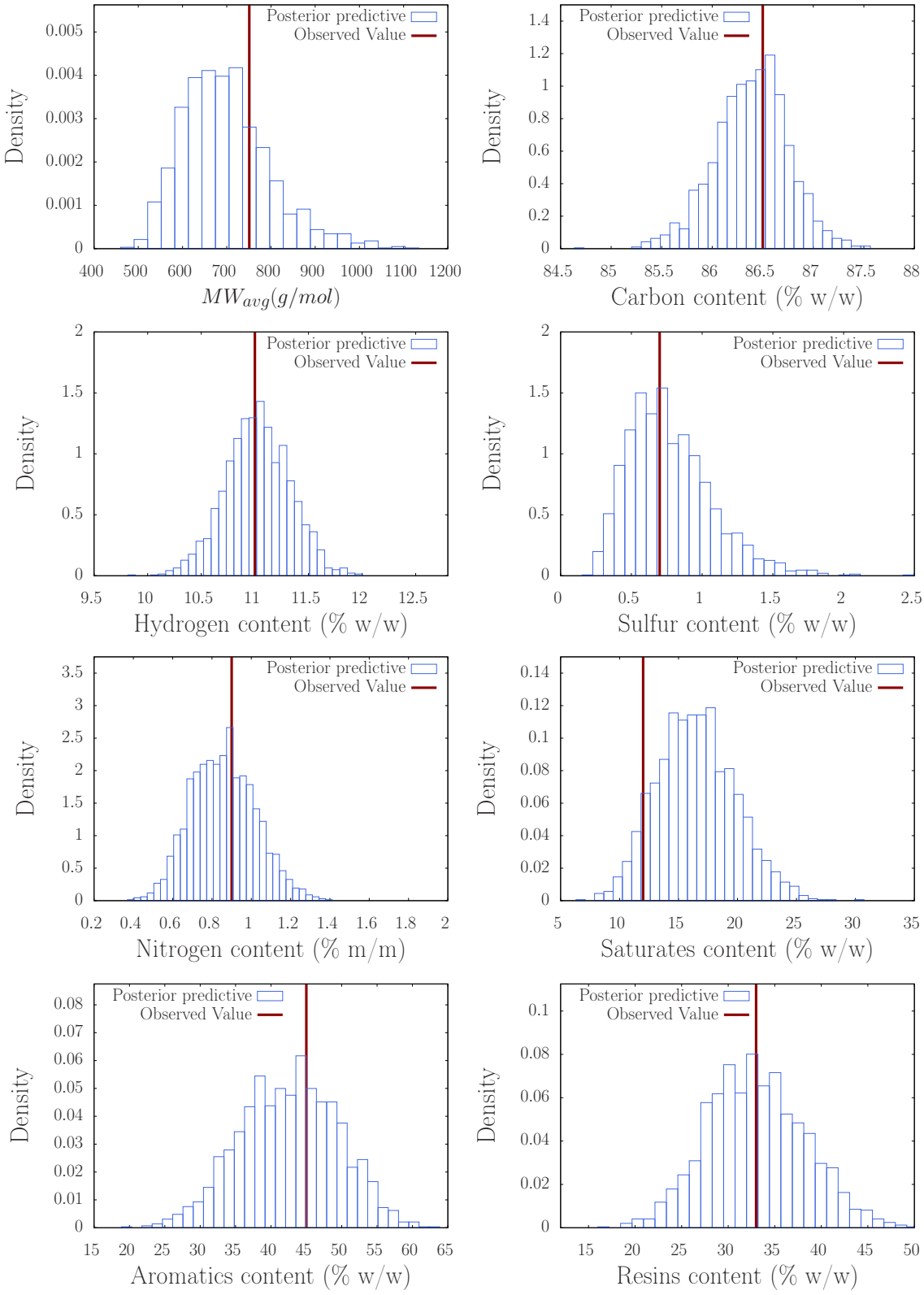


Figure 6.6: Posterior predictive distribution for the unconstrained properties. Each graph represents a different property. The observed value is included for comparison. The model can predict the new observed data. This shows the model ability to represent the molecular structures - Vacuum residue A.


Posterior predictive
Observed Value
Density
$MW_{avg}(g/mol)$
Carbon content (% w/w)
Hydrogen content (% w/w)
Sulfur content (% w/w)
Nitrogen content (% m/m)
Saturates content (% w/w)
Aromatics content (% w/w)
Resins content (% w/w)

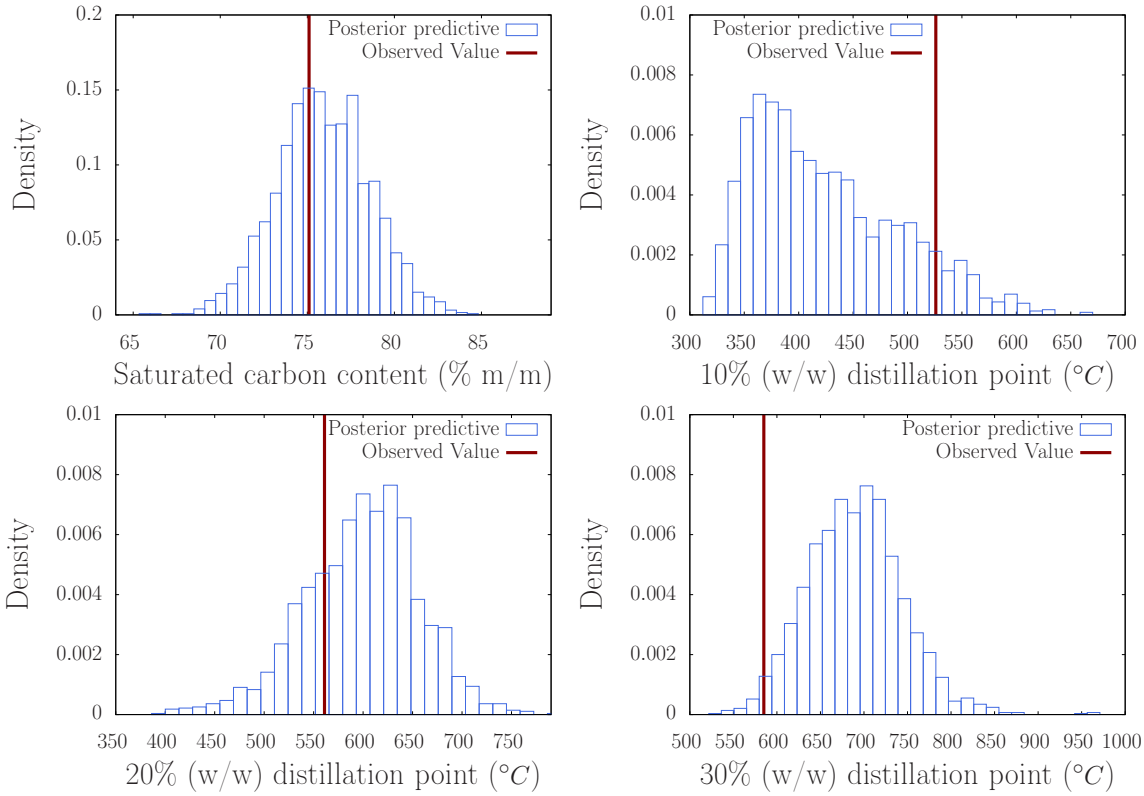
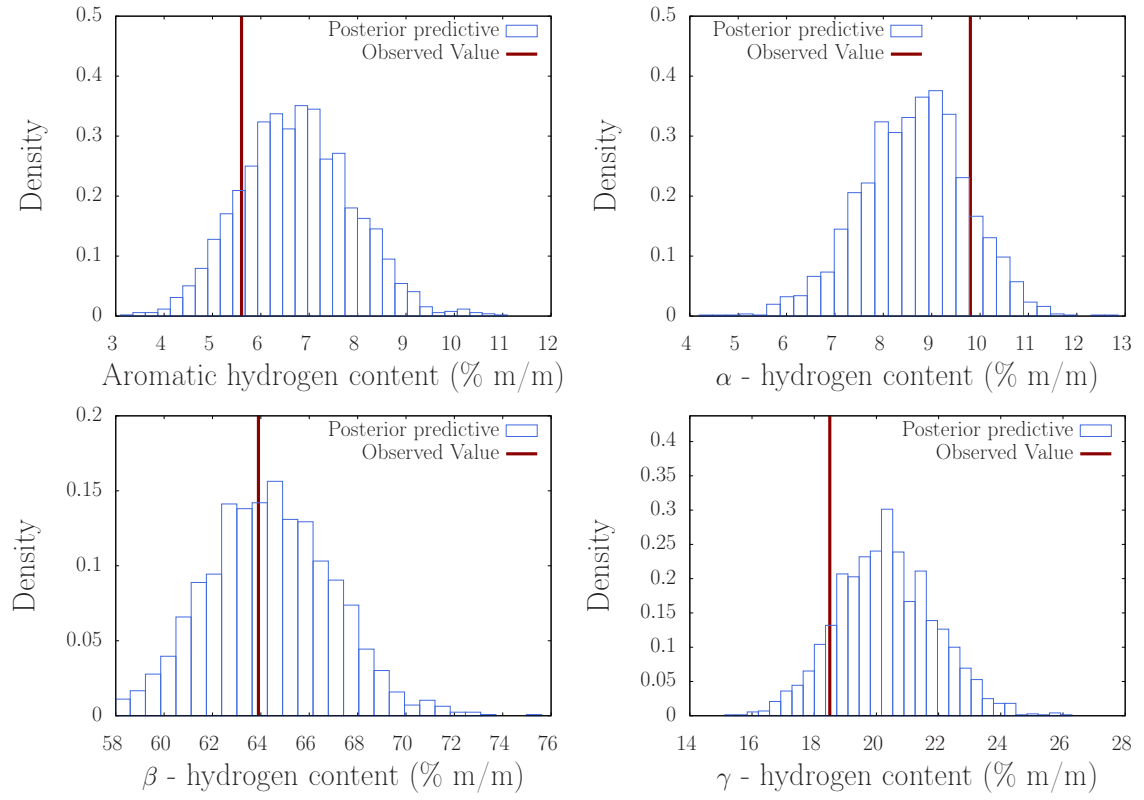


Figure 6.7: Posterior predictive distribution for the summary statistics (overall properties) used in the discrepancy function. Each graph represents a different property. The observed value is included for comparison. The model can replicate most of the observed data. An exception to the distillation curve - Vacuum residue B.



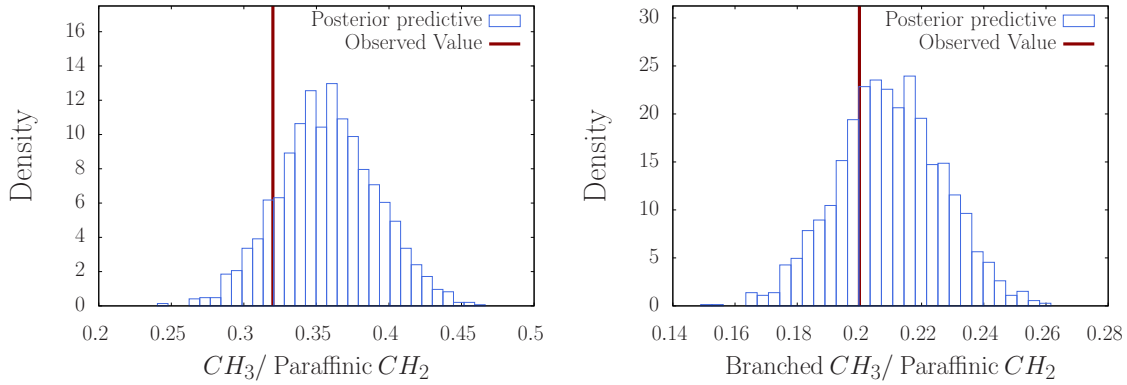


Figure 6.8: Posterior predictive distribution for the unconstrained properties. Each graph represents a different property. The observed value is included for comparison. The model can predict the new observed data. This shows the model ability to represent the molecular structures - Vacuum residue B.

6.2 Partioning around medoids and Reconstruction by entropy maximization

In the previous section, we showed the results from the stochastic reconstruction step in the algorithm developed here. The posterior distributions of the parameters controlling the molecular generation process were calculated for each of the vacuum residues studied. In this section, we will use these results to generate a molecular ensemble. We then choose a subset of representative molecules from the molecular ensemble using the partitioning around medoids algorithm. At last, we calculate the mixture composition using the reconstruction by entropy maximization algorithm.

To generate the initial molecular ensemble, we used the posterior mean as an estimator for the model parameters. We then ran the simulator to generate 5000 samples (molecules). With the generated molecules we can calculate the constraint matrix **g**, described in Section 5.6. We chose a total of 100 clusters (molecules) to represent the molecular ensemble.

In Tables 6.16 to 6.19, we show the predicted properties at each step of the algorithm. We can see that the proposed clustering technique is very effective in selecting the best representative molecules from the initial molecular ensemble. The predicted properties using the 100 clusters (molecules) are very close to the original mixture. The maximum entropy step brings the predicted properties to almost an exact match. Moreover, it corrects the discrepancies in the distillation curve.

Table 6.16: Observed and calculated properties after each step of the algorithm. Stochastic reconstruction (SR), Partitioning around medoids (PAM) and Reconstruction by entropy maximization (REM). Vacuum residue Ural.

Summary statistics	**Observed**	**After SR**	**After PAM**	**After REM**
MW_{avg} (g/mol)	727	755	756	727
Elemental Analysis				
Carbon content (% w/w)	85.5	85.1	85.2	85.5
Hydrogen content (% w/w)	10.6	10.7	10.5	10.6
Sulfur content (% w/w)	2.72	2.67	2.91	2.72
Nitrogen content (% w/w)	0.58	0.75	0.88	0.58
SARA fractions				
Saturates content (% w/w)	11.7	9.2	9.2	11.7
Aromatics content (% w/w)	46.1	48.0	48.0	46.1
Resins content(% w/w)	37.6	38.5	38.4	37.6
Nuclear magnetic resonance				
Saturated carbon content (% m/m)	72.8	73.2	70.8	72.8
Simulated distillation Temperature ($^{\circ}C$) at cumulative mass vaporization (%)				
10	520	581	576	518
20	550	686	709	565
30	574	746	754	578

Table 6.17: Observed and calculated properties after each step of the algorithm. Stochastic reconstruction (SR), Partitioning around medoids (PAM) and Reconstruction by entropy maximization (REM). Vacuum residue Maya.

Summary statistics	**Observed**	**After SR**	**After PAM**	**Afte REM**
MW_{avg} (g/mol)	764	742	746	764
Elemental Analysis				
Carbon content (% w/w)	85.2	84.7	84.4	85.2
Hydrogen content (% w/w)	10.1	10.4	10.3	10.1
Sulfur content (% w/w)	3.5	3.27	3.72	3.5
Nitrogen content (% w/w)	0.58	0.77	0.63	0.58
SARA fractions				
Saturates content (% w/w)	12.9	14.1	14.4	12.9
Aromatics content (% w/w)	38.7	37.9	37.9	38.7
Resins content(% w/w)	34.2	36.7	36.5	34.2
Nuclear magnetic resonance				
Saturated carbon content (% m/m)	69.5	70.5	69.7	69.5
Simulated distillation Temperature ($^{\circ}C$) at cumulative mass vaporization (%)				
10	520	570	567	520
20	550	668	674	554
30	574	732	752	574

Table 6.18: Observed and calculated properties after each step of the algorithm. Stochastic reconstruction (SR), Partitioning around medoids (PAM) and Reconstruction by entropy maximization (REM). Vacuum residue A.

Summary statistics	**Observed**	**After SR**	**After PAM**	**Afte REM**
MW_{avg} (g/mol)	718	676	675	718
Elemental Analysis				
Carbon content (% w/w)	86.7	86.2	86.4	86.7
Hydrogen content (% w/w)	11.4	11.4	11.2	11.4
Sulfur content (% w/w)	0.6	0.82	0.7	0.6
Nitrogen content (% w/w)	1.0	1.12	1.13	1.0
SARA fractions				
Saturates content (% w/w)	19.0	19.7	19.7	19.0
Aromatics content (% w/w)	40.0	43.7	43.7	40.0
Resins content(% w/w)	34.0	34.7	34.7	34.0
Nuclear magnetic resonance				
Saturated carbon content (% m/m)	77.6	79	80.4	77.6
Simulated distillation Temperature ($^{\circ}C$) at cumulative mass vaporization (%)				
10	525	368	478	518
20	555	593	619	553
30	576	683	700	601

Table 6.19: Observed and calculated properties after each step of the algorithm. Stochastic reconstruction (SR), Partitioning around medoids (PAM) and Reconstruction by entropy maximization (REM). Vacuum residue B.

Summary statistics	**Observed**	**After SR**	**After PAM**	**Afte REM**
MW_{avg} (g/mol)	751	681	683	751
Elemental Analysis				
Carbon content (% w/w)	86.5	86.4	86.6	86.5
Hydrogen content (% w/w)	11.0	11.1	11.2	11.0
Sulfur content (% w/w)	0.7	0.71	0.34	0.7
Nitrogen content (% w/w)	0.9	0.85	0.64	0.9
SARA fractions				
Saturates content (% w/w)	12.0	13.4	13.4	12.0
Aromatics content (% w/w)	45.0	47.5	47.5	45.0
Resins content(% w/w)	33.0	33.5	33.4	33.0
Nuclear magnetic resonance				
Saturated carbon content (% m/m)	75.1	75.9	78.1	75.1
Simulated distillation Temperature ($^{\circ}C$) at cumulative mass vaporization (%)				
10	526	412	403	524
20	561	600	585	566
30	585	684	672	586

Chapter 7

Conclusions

In this work, we developed a methodology to build molecules for heavy petroleum fractions based on limited experimental data. We called this methodology by molecular reconstruction. The algorithm is divided in three steps: stochastic reconstruction, partitioning around medoids and reconstruction by entropy maximization.

7.1 Stochastic reconstruction

We developed a novel algorithm combining the stochastic reconstruction methodologies and the structure-oriented lumping method for molecular representation and manipulation. We proposed an extension of the structure-oriented lumping method to increase molecular diversity.

We evaluated the stochastic reconstruction algorithm from a Bayesian perspective. The model can be classified as a simulator-based model. This allowed for a natural calculation of the uncertainty, both for the parameters and predictions.

The model was able to represent the vacuum residues studied in this work. Besides replicating the data from which it was trained, the model was also able to effectively predict new data.

7.2 Partitioning around medoids

The stochastic reconstruction algorithm can be seen as a Monte Carlo type procedure. For that, a large number of samples (molecules) is required to achieve a good representation of the population in question. That large number of molecules can be impeditive to some applications. Based on that, we developed a non-hierarchical clustering technique to select a subset of representative molecules from the initial molecular ensemble. Our method is based on the constraint matrix $\mathbf{g}$. Based on the calculated properties, the method was effective in choosing the most representative

molecules.

7.3 Reconstruction by entropy maximization

The stochastic reconstruction algorithm considers every built molecule to have the same importance in the mixture. In other words, it considers the mixture to be equimolar. Calculating mixture composition is a difficult task due to the lack of degrees of freedom. For that, we implemented a method called reconstruction by entropy maximization. The reconstruction by entropy maximization brings the calculated properties of the hypothetical mixtures to almost an exact match. Furthermore, it is able to correct the discrepancies in the distillation curve prediction observed in the previous steps.

7.4 Future work

- Evaluation of an optimal (or minimal) number of samples (molecules) to be taken from the stochastic reconstruction model. In this work, we used a fixed number of 5000. A systematic study of this variable would reduce the computational cost both for the parameter estimation step and for the subsequential uses of the generated molecular ensemble.

- Evaluation of an optimal number of clusters in the partitioning around medoids step of the algorithm. In this work, we used a fixed number of 100. Similar to the number of initial molecules, this study can potentially reduce even further the number of representative molecules. However, one may also conclude that the minimum number of clusters is higher than what was proposed here.

- Developing a methodology for uncertainty propagation from the stochastic reconstruction algorithm to the next steps. The usage of the Bayesian framework for parameter inference allowed a natural way to propagate uncertainty in the stochastic reconstruction step. However, in the reconstruction by entropy maximization step we only used the posterior distribution mean as an estimator of the parameters. Developing a methodology to handle uncertainties in the REM step is important for calculating prediction errors for the final models.

- Application of the generated molecular ensemble in modeling and simulation of refining processes. The main purpose of a detailed molecular characterization is a better representation of the refining processes, which might improve prediction and optimization of plant operations.

Bibliography

AHMAD, M. I., ZHANG, N., JOBSON, M., 2011, "Molecular components-based representation of petroleum fractions", *Chemical Engineering Research and Design*, v. 89, n. 4, pp. 410–420.

ALTGELT, K. H., BODUSZYNSKI, M. M., 1992, "Composition of heavy petroleums. 3. An improved boiling point-molecular weight relation", *Energy & fuels*, v. 6, n. 1, pp. 68–72.

ANCHEYTA, J., SÁNCHEZ, S., RODRÍGUEZ, M. A., 2005, "Kinetic modeling of hydrocracking of heavy oil fractions: A review", *Catalysis Today*, v. 109, n. 1-4, pp. 76–92.

AYE, M. M. S., ZHANG, N., 2005, "A novel methodology in transforming bulk properties of refining streams into molecular information", *Chemical engineering science*, v. 60, n. 23, pp. 6702–6717.

BEAUMONT, M. A., 2010, "Approximate Bayesian computation in evolution and ecology", *Annual review of ecology, evolution, and systematics*, v. 41, pp. 379–406.

BEAUMONT, M. A., ZHANG, W., BALDING, D. J., 2002, "Approximate Bayesian computation in population genetics", *Genetics*, v. 162, n. 4, pp. 2025–2035.

BEAUMONT, M. A., CORNUET, J.-M., MARIN, J.-M., et al., 2009, "Adaptive approximate Bayesian computation", *Biometrika*, v. 96, n. 4, pp. 983–990.

BETANCOURT, M., 2017, "A conceptual introduction to Hamiltonian Monte Carlo", *arXiv preprint arXiv:1701.02434*.

BLUM, M. G., FRANÇOIS, O., 2010, "Non-linear regression models for Approximate Bayesian Computation", *Statistics and Computing*, v. 20, n. 1, pp. 63–73.

BLUM, M. G., NUNES, M. A., PRANGLE, D., et al., 2013, "A comparative review of dimension reduction methods in approximate Bayesian computation", *Statistical Science*, v. 28, n. 2, pp. 189–208.

BODUSZYNSKI, M. M., 1987, "Composition of heavy petroleums. 1. Molecular weight, hydrogen deficiency, and heteroatom concentration as a function of atmospheric equivalent boiling point up to 1400. degree. F (760. degree. C)", *Energy & Fuels*, v. 1, n. 1, pp. 2–11.

BODUSZYNSKI, M. M., 1988, "Composition of heavy petroleums. 2. Molecular characterization", *Energy & Fuels*, v. 2, n. 5, pp. 597–613.

BODUSZYNSKI, M. M., ALTGELT, K. H., 1992, "Composition of heavy petroleums. 4. Significance of the extended atmospheric equivalent boiling point (AEBP) scale", *Energy & fuels*, v. 6, n. 1, pp. 72–76.

BROCHU, E., CORA, V. M., DE FREITAS, N., 2010, "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning", *arXiv preprint arXiv:1012.2599*.

CASELLA, G., BERGER, R. L., 2002, *Statistical inference*, v. 2. Duxbury Pacific Grove, CA.

CHARON-REVELLIN, N., DULOT, H., LÓPEZ-GARCÍA, C., et al., 2011, "Kinetic modeling of vacuum gas oil hydrotreatment using a molecular reconstruction approach", *Oil & Gas Science and Technology–Revue d'IFP Energies nouvelles*, v. 66, n. 3, pp. 479–490.

CHRISTENSEN, G., APELIAN, M. R., HICKEY, K. J., et al., 1999, "Future directions in modeling the FCC process: An emphasis on product quality", *Chemical Engineering Science*, v. 54, n. 13-14, pp. 2753–2764.

COKER, A. K., 2018, *Petroleum Refining Design and Applications Handbook*. John Wiley & Sons.

CSILLÉRY, K., BLUM, M. G., GAGGIOTTI, O. E., et al., 2010, "Approximate Bayesian computation (ABC) in practice", *Trends in ecology & evolution*, v. 25, n. 7, pp. 410–418.

DE OLIVEIRA, L. P., VAZQUEZ, A. T., VERSTRAETE, J. J., et al., 2013, "Molecular Reconstruction of Petroleum Fractions: Application to Vacuum Residues from Different Origins", *Energy & Fuels*, v. 27, n. 7,

pp. 3622–3641. doi: 10.1021/ef300768u. Disponível em: <https://doi.org/10.1021/ef300768u>.

DE OLIVEIRA, L. P., VERSTRAETE, J. J., KOLB, M., 2012, "A Monte Carlo modeling methodology for the simulation of hydrotreating processes", *Chemical engineering journal*, v. 207, pp. 94–102.

DE OLIVEIRA, LUÍS P., HUDEBINE, DAMIEN, GUILLAUME, DENIS, et al., 2016, "A Review of Kinetic Modeling Methodologies for Complex Processes", *Oil Gas Sci. Technol. – Rev. IFP Energies nouvelles*, v. 71, n. 3, pp. 45. doi: 10.2516/ogst/2016011. Disponível em: <https://doi.org/10.2516/ogst/2016011>.

DEL MORAL, P., DOUCET, A., JASRA, A., 2012, "An adaptive sequential Monte Carlo method for approximate Bayesian computation", *Statistics and Computing*, v. 22, n. 5, pp. 1009–1020.

DENIZ, C. U., YASAR, M., KLEIN, M. T., 2017a, "Stochastic Reconstruction of Complex Heavy Oil Molecules Using an Artificial Neural Network", *Energy & Fuels*, v. 31, n. 11, pp. 11932–11938. doi: 10.1021/acs.energyfuels.7b02311. Disponível em: <https://doi.org/10.1021/acs.energyfuels.7b02311>.

DENIZ, C. U., YASAR, M., KLEIN, M. T., 2017b, "A New Extended Structural Parameter Set for Stochastic Molecular Reconstruction: Application to Asphaltenes", *Energy & Fuels*, v. 31, n. 8, pp. 7919–7931. doi: 10.1021/acs.energyfuels.7b01006. Disponível em: <https://doi.org/10.1021/acs.energyfuels.7b01006>.

DUTRIEZ, T., COURTIADE, M., THIÉBAUT, D., et al., 2010, "Improved hydrocarbons analysis of heavy petroleum fractions by high temperature comprehensive two-dimensional gas chromatography", *Fuel*, v. 89, n. 9, pp. 2338–2345.

DUTRIEZ, T., BORRAS, J., COURTIADE, M., et al., 2011, "Challenge in the speciation of nitrogen-containing compounds in heavy petroleum fractions by high temperature comprehensive two-dimensional gas chromatography", *Journal of Chromatography A*, v. 1218, n. 21, pp. 3190–3199.

DUTTA, R., KASKI, S., LINTUSAARI, J., et al., 2016, "Fundamentals and Recent Developments in Approximate Bayesian Computation", *Systematic Biology*, v. 66, n. 1.

FAN, T., WANG, J., BUCKLEY, J. S., et al., 2002, "Evaluating crude oils by SARA analysis". In: *SPE/DOE improved oil recovery symposium*. Society of Petroleum Engineers.

FEARNHEAD, P., PRANGLE, D., 2012, "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, v. 74, n. 3, pp. 419–474.

FLORY, P. J., 1952, "Molecular size distribution in three dimensional polymers. VI. Branched polymers containing A—R—Bf-1 type units", *Journal of the American Chemical Society*, v. 74, n. 11, pp. 2718–2723.

GELMAN, A., CARLIN, J. B., STERN, H. S., et al., 2014, *Bayesian data analysis*, v. 2. CRC press Boca Raton, FL.

GOMEZ-PRADO, J., ZHANG, N., THEODOROPOULOS, C., 2008, "Characterisation of heavy petroleum fractions using modified molecular-type homologous series (MTHS) representation", *Energy*, v. 33, n. 6, pp. 974–987.

GRAY, M. R., 2003, "Consistency of asphaltene chemical structures with pyrolysis and coking behavior", *Energy & Fuels*, v. 17, n. 6, pp. 1566–1569.

GRAY, M. R., MCCAFFREY, W. C., 2002, "Role of chain reactions and olefin formation in cracking, hydroconversion, and coking of petroleum and bitumen fractions", *Energy & fuels*, v. 16, n. 3, pp. 756–766.

GRAY, R. M., 1994, *Upgrading petroleum residues and heavy oils*. CRC press.

GUTMANN, M. U., CORANDER, J., 2016, "Bayesian optimization for likelihood-free inference of simulator-based statistical models", *The Journal of Machine Learning Research*, v. 17, n. 1, pp. 4256–4302.

HASAN, M. U., ALI, M. F., BUKHARI, A., 1983, "Structural characterization of Saudi Arabian heavy crude oil by nmr spectroscopy", *Fuel*, v. 62, n. 5, pp. 518 – 523.

HICKERSON, M. J., STAHL, E. A., LESSIOS, H. A., 2006, "Test for simultaneous divergence using approximate Bayesian computation", *Evolution*, v. 60, n. 12, pp. 2435–2453.

HIRSCH, E., ALTGELT, K. H., 1970, "Integrated structural analysis. Method for the determination of average structural parameters of petroleum heavy ends", *Analytical Chemistry*, v. 42, n. 12, pp. 1330–1339.

HOFFMAN, M. D., GELMAN, A., 2014, "The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research*, v. 15, n. 1, pp. 1593–1623.

HU, S., TOWLER, G., ZHU, X., 2002, "Combine molecular modeling with optimization to stretch refinery operation", *Industrial & engineering chemistry research*, v. 41, n. 4, pp. 825–841.

HUDEBINE, D., VERSTRAETE, J. J., 2011, "Reconstruction of petroleum feedstocks by entropy maximization. Application to FCC gasolines", *Oil & Gas Science and Technology–Revue d'IFP Energies nouvelles*, v. 66, n. 3, pp. 437–460.

HUDEBINE, D., VERA, C., WAHL, F., et al., 2002, "Molecular representation of hydrocarbon mixtures from overall petroleum analyses". In: *AIChE Spring Meeting*, pp. 10–14.

HUDEBINE, D., VERSTRAETE, J. J., 2004, "Molecular reconstruction of LCO gasoils from overall petroleum analyses", *Chemical Engineering Science*, v. 59, n. 22-23, pp. 4755–4763.

JAFFE, S. B., FREUND, H., OLMSTEAD, W. N., 2005, "Extension of structure-oriented lumping to vacuum residua", *Industrial & engineering chemistry research*, v. 44, n. 26, pp. 9840–9852.

JÄRVENPÄÄ, M., GUTMANN, M. U., PLESKA, A., et al., 2019, "Efficient acquisition rules for model-based approximate Bayesian computation", *Bayesian Analysis*.

JAYNES, E. T., 1957, "Information theory and statistical mechanics", *Physical review*, v. 106, n. 4, pp. 620.

KLEIN, M. T., HOU, G., BERTOLACINI, R., et al., 2005, *Molecular modeling in heavy hydrocarbon conversions*. CRC Press.

LUMPKIN, H., 1956, "Determination of saturated hydrocarbons in heavy petroleum fractions by mass spectrometry", *Analytical Chemistry*, v. 28, n. 12, pp. 1946–1948.

MARJORAM, P., MOLITOR, J., PLAGNOL, V., et al., 2003, "Markov chain Monte Carlo without likelihoods", *Proceedings of the National Academy of Sciences*, v. 100, n. 26, pp. 15324–15328.

MCKENNA, A. M., BLAKNEY, G. T., XIAN, F., et al., 2010a, "Heavy petroleum composition. 2. Progression of the Boduszynski model to the limit of distillation by ultrahigh-resolution FT-ICR mass spectrometry", *Energy & Fuels*, v. 24, n. 5, pp. 2939–2946.

MCKENNA, A. M., PURCELL, J. M., RODGERS, R. P., et al., 2010b, "Heavy petroleum composition. 1. Exhaustive compositional analysis of Athabasca bitumen HVGO distillates by Fourier transform ion cyclotron resonance mass spectrometry: A definitive test of the Boduszynski model", *Energy & Fuels*, v. 24, n. 5, pp. 2929–2938.

MCKENNA, A. M., DONALD, L. J., FITZSIMMONS, J. E., et al., 2013a, "Heavy petroleum composition. 3. Asphaltene aggregation", *Energy & fuels*, v. 27, n. 3, pp. 1246–1256.

MCKENNA, A. M., MARSHALL, A. G., RODGERS, R. P., 2013b, "Heavy petroleum composition. 4. Asphaltene compositional space", *Energy & Fuels*, v. 27, n. 3, pp. 1257–1267.

MEAD, W. L., 1968, "Field ionization mass spectrometry of heavy petroleum fractions. Waxes", *Analytical Chemistry*, v. 40, n. 4, pp. 743–747.

NEUROCK, M., LIBANATI, C., NIGAM, A., et al., 1990, "Monte Carlo simulation of complex reaction systems: Molecular structure and reactivity in modelling heavy oils", *Chemical Engineering Science*, v. 45, n. 8, pp. 2083–2088.

NEUROCK, M., NIGAM, A., TRAUTH, D., et al., 1994, "Molecular representation of complex hydrocarbon feedstocks through efficient characterization and stochastic algorithms", *Chemical engineering science*, v. 49, n. 24, pp. 4153–4177.

PARK, H.-S., JUN, C.-H., 2009, "A simple and fast algorithm for K-medoids clustering", *Expert systems with applications*.

PENG, B., 1999, *Molecular modelling of petroleum processes*. Tese de Doutorado, UMIST.

PODGORSKI, D. C., CORILO, Y. E., NYADONG, L., et al., 2013, "Heavy petroleum composition. 5. Compositional and structural continuum of petroleum revealed", *Energy & Fuels*, v. 27, n. 3, pp. 1268–1276.

PRESSÉ, S., GHOSH, K., LEE, J., et al., 2013, "Principles of maximum entropy and maximum caliber in statistical physics", *Reviews of Modern Physics*, v. 85, n. 3, pp. 1115.

PYL, S. P., HOU, Z., VAN GEEM, K. M., et al., 2011, "Modeling the composition of crude oil fractions using constrained homologous series", *Industrial & Engineering Chemistry Research*, v. 50, n. 18, pp. 10850–10858.

QIAN, K., ROBBINS, W. K., HUGHEY, C. A., et al., 2001a, "Resolution and identification of elemental compositions for more than 3000 crude acids in heavy petroleum by negative-ion microelectrospray high-field Fourier transform ion cyclotron resonance mass spectrometry", *Energy & Fuels*, v. 15, n. 6, pp. 1505–1511.

QIAN, K., RODGERS, R. P., HENDRICKSON, C. L., et al., 2001b, "Reading chemical fine print: Resolution and identification of 3000 nitrogen-containing aromatic compounds from a single electrospray ionization Fourier transform ion cyclotron resonance mass spectrum of heavy petroleum crude oil", *Energy & Fuels*, v. 15, n. 2, pp. 492–498.

QIAN, K., EDWARDS, K. E., SISKIN, M., et al., 2007, "Desorption and ionization of heavy petroleum molecules and measurement of molecular weight distributions", *Energy & fuels*, v. 21, n. 2, pp. 1042–1047.

QUANN, R. J., 1998, "Modeling the chemistry of complex petroleum mixtures." *Environmental Health Perspectives*, v. 106, n. Suppl 6, pp. 1441.

QUANN, R. J., JAFFE, S. B., 1992, "Structure-oriented lumping: describing the chemistry of complex hydrocarbon mixtures", *Industrial & engineering chemistry research*, v. 31, n. 11, pp. 2483–2497.

QUANN, R., JAFFE, S., 1996, "Building useful models of complex reaction systems in petroleum refining", *Chemical Engineering Science*, v. 51, n. 10, pp. 1615–1635.

RIAZI, M., 2005, *Characterization and properties of petroleum fractions*. ASTM international.

ROBERT, C. P., CORNUET, J.-M., MARIN, J.-M., et al., 2011, "Lack of confidence in approximate Bayesian computation model choice", *Proceedings of the National Academy of Sciences*, v. 108, n. 37, pp. 15112–15117.

ROSE, K., FRANCISCO, M., 1987, "Characterization of acidic heteroatoms in heavy petroleum fractions by phase-transfer methylation and NMR spectroscopy", *Energy & fuels*, v. 1, n. 3, pp. 233–239.

ROUSSIS, S. G., PROULX, R., 2002, "Molecular weight distributions of heavy aromatic petroleum fractions by Ag+ electrospray ionization mass spectrometry", *Analytical chemistry*, v. 74, n. 6, pp. 1408–1414.

ROUSSIS, S. G., PROULX, R., 2004, "Probing the molecular weight distributions of non-boiling petroleum fractions by Ag+ electrospray ionization mass spectrometry", *Rapid communications in mass spectrometry*, v. 18, n. 15, pp. 1761–1775.

SAWATZKY, H., GEORGE, A. E., SMILEY, G. T., et al., 1976, "Hydrocarbon-type separation of heavy petroleum fractions", *Fuel*, v. 55, n. 1, pp. 16–20.

SHANNON, C. E., 1948, "A mathematical theory of communication", *Bell system technical journal*, v. 27, n. 3, pp. 379–423.

SPEIGHT, J. G., 2014, *The chemistry and technology of petroleum*. CRC press.

SPEIGHT, J. G., OZUM, B., 2001, *Petroleum refining processes*. CRC Press.

TIAN, L., WANG, J., SHEN, B., et al., 2010, "Building a kinetic model for steam cracking by the method of structure-oriented lumping", *Energy & fuels*, v. 24, n. 8, pp. 4380–4386.

TIAN, L., SHEN, B., LIU, J., 2012a, "A delayed coking model built using the structure-oriented lumping method", *Energy & Fuels*, v. 26, n. 3, pp. 1715–1724.

TIAN, L., SHEN, B., LIU, J., 2012b, "Building and application of delayed coking structure-oriented lumping model", *Industrial & Engineering Chemistry Research*, v. 51, n. 10, pp. 3923–3931.

TONI, T., WELCH, D., STRELKOWA, N., et al., 2009, "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems", *Journal of the Royal Society Interface*, v. 6, n. 31, pp. 187–202.

TRAUTH, D. M., STARK, S. M., PETTI, T. F., et al., 1994, "Representation of the molecular structure of petroleum resid through characterization and Monte Carlo modeling", *Energy & fuels*, v. 8, n. 3, pp. 576–580.

TRESTIANU, S., ZILIOLI, G., SIRONI, A., et al., 1985, "Automatic simulated distillation of heavy petroleum fractions up to 800 C TBP by capillary gas chromatography. Part I: Possibilities and limits of the method", *Journal of Separation Science*, v. 8, n. 11, pp. 771–781.

VAN GEEM, K. M., HUDEBINE, D., REYNIERS, M. F., et al., 2007, "Molecular reconstruction of naphtha steam cracking feedstocks based on commercial indices", *Computers & chemical engineering*, v. 31, n. 9, pp. 1020–1034.

VASSILIOU, M. S., 2018, *Historical dictionary of the petroleum industry*. Rowman & Littlefield.

VERSTRAETE, J. J., REVELLIN, N., DULOT, H., et al., 2004, "Molecular reconstruction of vacuum gasoils", *Prepr. Pap.-Am. Chem. Soc., Div. Fuel Chem*, v. 49, n. 1, pp. 20.

VERSTRAETE, J., SCHNONGS, P., DULOT, H., et al., 2010, "Molecular reconstruction of heavy petroleum residue fractions", *Chemical Engineering Science*, v. 65, n. 1, pp. 304 – 312. ISSN: 0009-2509. doi: https://doi.org/10.1016/j.ces.2009.08.033. Disponível em: <`http://www.sciencedirect.com/science/article/pii/S000925090900565X`>.
20th International Symposium in Chemical Reaction Engineering—Green Chemical Reaction Engineering for a Sustainable Future.

WALDO, G. S., CARLSON, R. M., MOLDOWAN, J. M., et al., 1991, "Sulfur speciation in heavy petroleums: Information from X-ray absorption near-edge structure", *Geochimica et Cosmochimica Acta*, v. 55, n. 3, pp. 801–814.

WEGMANN, D., LEUENBERGER, C., EXCOFFIER, L., 2009, "Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood", *Genetics*, v. 182, n. 4, pp. 1207–1218.

WEI, W., BENNETT, C. A., TANAKA, R., et al., 2008, "Computer aided kinetic modeling with KMT and KME", *Fuel Processing Technology*, v. 89, n. 4, pp. 350–363.

WILKINSON, R. D., 2013, "Approximate Bayesian computation (ABC) gives exact results under the assumption of model error", *Statistical applications in genetics and molecular biology*, v. 12, n. 2, pp. 129–141.

WU, Y., ZHANG, N., 2010, "Molecular characterization of gasoline and diesel streams", *Industrial & Engineering Chemistry Research*, v. 49, n. 24, pp. 12773–12782.

YANG, B., ZHOU, X., CHEN, C., et al., 2008, "Molecule simulation for the secondary reactions of fluid catalytic cracking gasoline by the method of structure oriented lumping combined with Monte Carlo", *Industrial & Engineering Chemistry Research*, v. 47, n. 14, pp. 4648–4657.