Université de Lille - Sciences et Technologies
École Doctorale des Sciences de la Matière, du Rayonnement et de l'Environnement

Investigation of descriptors for the understanding and prediction of fluid oxidation stability by machine learning

Thèse de Doctorat

préparée et soutenue publiquement par

Adrián Venegas Reynoso

le 27 mars 2025

pour obtenir le grade de

Docteur

en

Chimie théorique, physique, analytique

Composition du jury:

Président du jury et rapporteur:	
M. Pierre-Alexandre GLAUDE	Directeur de recherche, Université de Lorraine, France
Rapporteur:	
M. Jean-Michel ROGER	Directeur de recherche, Université de Montpellier, France
Examinateurs:	
M. Alexandre VARNEK	Professeur, Université de Strasbourg, France
Mme. Laurie STARCK	Docteure, Societé Chevron Oronite, France
Directeur de thèse:	
M. Ludovic DUPONCHEL	Professeur, Université de Lille, France
Co-encadrante de thèse:	
Mme. Lucia GIARRACCA-MEHL	Docteure, IFP Energies nouvelles, France
Invités:	
M. Benoît CRETON	Docteur, IFP Energies nouvelles, France
Mme. Marion LACOUE-NEGRE	Docteure, IFP Energies nouvelles, France

Université de Lille - Sciences et Technologies
École Doctorale des Sciences de la Matière, du Rayonnement et de l'Environnement

Investigation de descripteurs pour la compréhension et la prédiction de la stabilité à l'oxydation des fluides par apprentissage automatique

Thèse de Doctorat

préparée et soutenue publiquement par

Adrián Venegas Reynoso

le 27 mars 2025

pour obtenir le grade de

Docteur

en

Chimie théorique, physique, analytique

Composition du jury:

Président du jury et rapporteur:
M. Pierre-Alexandre GLAUDE — Directeur de recherche, Université de Lorraine, France

Rapporteur:
M. Jean-Michel ROGER — Directeur de recherche, Université de Montpellier, France

Examinateurs:
M. Alexandre VARNEK — Professeur, Université de Strasbourg, France
Mme. Laurie STARCK — Docteure, Societé Chevron Oronite, France

Directeur de thèse:
M. Ludovic DUPONCHEL — Professeur, Université de Lille, France

Co-encadrante de thèse:
Mme. Lucia GIARRACCA-MEHL — Docteure, IFP Energies nouvelles, France

Invités:
M. Benoît CRETON — Docteur, IFP Energies nouvelles, France
Mme. Marion LACOUE-NEGRE — Docteure, IFP Energies nouvelles, France

Abstract

The aviation sector contributes approximately 4% of global CO2 emissions and remains one of the most significant industries worldwide. However, reducing emissions from aviation presents substantial challenges. Electrification is constrained by the low energy density of current batteries, while the adoption of cryogenic hydrogen is limited by the development of lightweight storage tanks.

Conversely, significant research efforts have focused on the development and certification of synthetic fuels derived from renewable feedstocks, commonly referred to as Sustainable Aviation Fuel (SAF). Although both conventional jet fuels and SAF are generally expected to maintain their properties over time, exposure to pollutants and thermal stress can result in oxidation. Oxidation degrades fuel quality, reduces system efficiency, and can even cause operational failures.

Kinetic modeling has been a key tool for studying oxidation phenomena, but its mechanisms are typically developed for the gas phase. Extending these models to account for solvent effects in the liquid phase is particularly challenging, especially for complex mixtures like fuels.

To address these challenges, this thesis employs data-driven modeling techniques to study oxidation phenomena, including Quantitative Structure-Property Relationships (QSPRs) and Near-Infrared (NIR) spectroscopy-based modeling. A notable limitation in the literature is the scarcity of oxidation data for pure hydrocarbons, prompting the development of a new database in this work.

The Rapid Small-Scale Oxidation Test (RSSOT), also known as the PetroOxy or RapidOxy test, was employed to measure the Induction Period (IP), a parameter that quantifies the time required for a sample to react with a given amount of oxygen. This study analyzed 95 hydrocarbons across a temperature range of 40°C to 160°C, identifying reactivity trends based on molecular features.

Using these trends, a predictive QSPR model for the Induction Period was developed, testing multiple machine learning algorithms, including Support Vector Machine with a radial basis function (RBF) kernel, XGBoost Tree, and XGBoost Linear. The resulting semi-quantitative model serves as a useful tool for screening potential fuel candidates. For NIR spectroscopy-based modeling of the Induction Period, the study employed Support Vector Machine with an

RBF kernel. However, this approach yielded a model with low accuracy. Based on the observed reactivity trends, the limitations of NIR spectroscopy for predicting the Induction Period are discussed. Finally, this work proposes Nuclear Magnetic Resonance (NMR) spectroscopy as a more suitable technique for this purpose and presents modeling results based on simulated NMR data.

Résumé

Le secteur de l'aviation contribue à environ 4% des émissions mondiales de CO2 et demeure l'une des industries les plus importantes à l'échelle mondiale. Cependant, réduire les émissions dans ce secteur demeure un défi considérable. En effet, l'électrification est limitée par la faible densité énergétique des batteries actuelles, tandis que l'adoption de l'hydrogène cryogénique est freinée par le développement de réservoirs de stockage légers. En revanche, des efforts de recherche significatifs se sont concentrés sur le développement et la certification de carburants synthétiques issus de matières premières renouvelables, communément appelés carburants durables d'aviation (SAF, pour Sustainable Aviation Fuels).

Bien que les carburants conventionnels pour avions et les SAF soient généralement supposés conserver leurs propriétés au fil du temps, leur exposition aux polluants et au stress thermique peut entraîner une oxydation. L'oxydation dégrade la qualité des carburants, réduit l'efficacité des systèmes et peut même provoquer des défaillances opérationnelles. La modélisation cinétique a été un outil clé pour étudier les phénomènes d'oxydation, mais ses mécanismes sont généralement développés pour la phase gazeuse. Étendre ces modèles pour tenir compte des effets du solvant en phase liquide est particulièrement difficile, notamment pour des mélanges complexes comme les carburants.

Pour relever ces défis, cette thèse utilise des techniques de modélisation basées sur les données pour étudier les phénomènes d'oxydation, notamment les relations quantitatives structure-propriété (QSPR) et la modélisation basée sur la spectroscopie proche infrarouge (NIR). Une limitation notable observée dans la littérature est la rareté des données d'oxydation pour les hydrocarbures purs, ce qui a tout d'abord conduit au développement d'une nouvelle base de données dans ce travail de thèse. Le test d'oxydation rapide à petite échelle (RSSOT), également connu sous le nom de test PetroOxy ou RapidOxy, a été utilisé pour mesurer la période d'induction (IP), un paramètre qui quantifie le temps nécessaire pour qu'un échantillon réagisse avec une quantité donnée d'oxygène.

Cette étude a portée sur l'analyse de 95 hydrocarbures sur une plage de température allant de 40 °C à 160 °C, identifiant des tendances de réactivité basées sur les caractéristiques moléculaires. À partir de ces tendances, un modèle prédictif QSPR pour la période d'induction a été développé, en testant plusieurs algorithmes d'apprentissage automatique, notamment la machine à vecteurs de support (SVM) avec un noyau à base radiale (RBF), XGBoost Tree et

XGBoost Linear. Le modèle semi-quantitatif résultant constitue un outil utile pour le criblage des candidats potentiels pour les carburants.

Pour la modélisation basée sur la spectroscopie NIR de la période d'induction, l'étude a utilisé une machine à vecteurs de support avec un noyau RBF. Cependant, cette approche a donné un modèle de faible précision. Sur la base des tendances de réactivité observées, les limites de la spectroscopie NIR pour prédire la période d'induction sont discutées. Enfin, ce travail propose la spectroscopie par résonance magnétique nucléaire (RMN) comme une technique plus appropriée à cette fin et présente des résultats de modélisation basés sur des données RMN simulées.

Acknowledgements

I considered writing a verbatim transcript of the acknowledgement speech I gave after my defense. However, while aiming to remain truthful to what I said, I also took the opportunity to expand on the original speech.

The defense of my PhD marked a significant milestone in both my academic and personal development. Nevertheless, this achievement would not have been possible without the support and help of many people, to whom I owe my deepest gratitude.

First, I thank the members of the jury for accepting to review my work, for attending my defense—whether in person or remotely—and for their insightful and constructive feedback. I am also deeply grateful to all my supervisors for their close involvement in my work. I begin by thanking Lucia Giarracca for giving me the opportunity to work on this fascinating project and for seeing qualities in me that made her believe I was the right person for it. I also sincerely thank my co-supervisors, Benoît Creton and Marion Lacuoe-Negre, who, despite "only" being co-supervisors, were consistently engaged with the project and my progress, offering valuable guidance—both academically and personally. I am grateful to Cyril Ruckebusch, whose insight helped shape the early stages of this project, even though he later had to step down from his role as co-director.

I especially want to thank my director, Ludovic Duponchel. As I jokingly mentioned during my defense, Ludovic, I must admit that presenting my results to you often made me nervous and left me stuttering. Yet your rigor and high standards have helped shape me into a far better scientist. For that, I can only express my profound gratitude and respect.

After some introspection, I feel I must also acknowledge several professors who had a lasting impact on my academic formation. In particular, I would like to thank Ramón Solorio, who ignited my passion for chemistry fifteen years ago. I also want to thank Roberto Flores and Gilberto Velázquez, who played pivotal roles in my development as a chemist during my undergraduate studies.

I am thankful to my IFPEN colleagues-turned-friends—Nathalie Brassart, Mickaël Matrat, and Boyang Xu—for their collaboration and friendship. I also want to thank my fellow PhD students and dear friends: Adan, Carlos, Alejandra, Rosa, Laura, Andrea, Karla, Ryma, and Fabiola, for all the laughs, drinks, and joyful moments. Additionally, I want to express my

heartfelt thanks to some of my closest friends—Mónica, Carlos Andrés, Michał, Lucas, and Luca—for their continued friendship and support.

I thought the best way to thank my family was to do so in my native language: *A mi familia; mis padres y mi hermano, gracias por su apoyo y amor, no sólo durante estos tres años, sino durante toda mi vida. Gracias por inspirarme a siempre dar lo mejor de mí. Este logro está dedicado a ustedes.*

Lastly, I want to thank my fiancée, Aleksandra, for her love, unconditional support, and, for lack of a better term, putting up with me during all these years.

Contents

List of Figures

List of Tables

List of Acronyms

AD Applicability Domain.
ALS Asymmetric Least Squares.
ANN Artificial Neural Network.
ASTM American Society for Testing and Materials.
ATJ-SKA Alcohol-To-Jet Synthetic Paraffinic Kerosene with Aromatics.
ATJ-SPK Alcohol-To-Jet Synthetic Paraffinic Kerosene.

BCI Barnard Chemistry Information.
BDE Bond-Dissociation Energy.
BHT Butylated Hydroxytoluene.
BIC Biofuels Iso-Conversion.

CHJ Catalytic Hydrothermolysis Jet.
COSMO COnductor-like Screening Calculation MOdel.
CV Cross-Validation.

DA Data Augmentation.
DARTMS Direct Analysis in Real Time Mass Spectrometry.
DFT Density Functional Theory.
DILSAAF Single Reactor HEFA (Drop-in Liquid Sustainable Aviation and Automotive Fuel).
DTGS Deuterated Triglycine Sulfate.

ECFPs Extended Connectivity FingerPrints.
EN European Standard.
EPR Electron Paramagnetic Resonance.
Equation of State Equation of State.
ESR Electron Spin Resonance.

FAME Fatty Acid Methyl Ester.
FCFPs Functional-Class FingerPrints.
FGCD Functional Group Count Descriptors.
FIR Far-Infrared.
FT Fischer-Tropsch.
FTICRMS Fourier Transform Ion Cyclotron Resonance Mass Spectrometry.
FTIRM Fourier Transform Infrared Microscopy.

GCM-UOB Group Contribution Method of the University of Birmingham.
GCxGC Comprehensive two-dimensional gas chromatography.
GDP Gross Domestic Product.
GETAWAY GEometry, Topology, and Atom-Weights AssemblY.
GHG Greenhouse Gases.

HC-HEFA SPK Synthesized Paraffinic Kerosene from bio-derived Hydroprocessed Hydrocarbons, Esters and Fatty Acids.
HEFA SPK Synthesized Paraffinic Kerosene from Hydroprocessed Esters and Fatty Acids.
HOMO Highest Occupied Molecular Orbital.

IATA International Air Transport Association.

ICAO International Civil Aviation Organization.
IFF Interesting Feature Finder.
IH2 Integrated Hydropyrolysis and Hydroconversion.
InChI International Chemical Identifier.
IP Induction Period.
IR Infrared.
IRLS Iteratively Reweighted Least Squares.
ISO International Organization for Standardization.
IUPAC International Union of Pure and Applied Chemistry.
IV Iodine Value.

JFTOT Jet Fuel Thermal Oxidation Stability Test.
JP Jet Propellant.

LASSO Least Absolute Shrinkage and Selection Operator.
LOO Leave-One-Out.
LSER Linear Solvation Energy Relationship.
LUMO Lowest Unoccupied Molecular Orbital.

M-QSPR Mean Quantitative Structure-Property Relationship.
MACCS Molecular ACCess System.
MCR Multivariate Curve Resolution.
MIR Mid-Infrared.
ML Machine-Learning.
MoRSE 3D-Molecular Representation of Structures based on Electron diffraction.
MSC Multiplicative Scatter Correction.
MTJ Methanol-to-Jet.
MWA Modified Weighted Average.

NIR Near-Infrared.
NIRS Near-Infrared Spectroscopy.
NMR Nuclear Magnetic Resonance.

OSI Oil Stability Index.

PC Principal Component.
PCA Principal Component Analysis.
PCM Polarizable Continuum Model.
PLS Partial Least Squares.
PLSR Partial Least Squares Regression.
PM Particulate Matters.
Polycyclic Aromatic Hydrocarbon PAH.

QCM Quartz Crystal Microbalance.
QSAR Quantitative Structure-Activity Relationship.
QSPR Quantitative Structure-Property Relationship.
QSTR Quantitative Structure-Toxicity Relationship.

RBF Radial Basis Function.
ReOIL Pyrolysis of non-recyclable plastics.
RMSE Root Mean Square Error.
RMSEC Root Mean Square Error of the Calibration.
RMSECV Root Mean Square Error of the Cross-Validation.
RMSEP Root Mean Square Error of the Prediction.
RSSOT Rapid Small Scale Oxidation Test.

SAF Sustainable Aviation Fuel.
SAK Synthesized Aromatic Kerosene.
SEM Scanning Electron Microscopy.
SG Savitzky-Golay.
SHAP SHapley Additive exPlanations.
SIP Synthesized Iso-paraffins.
SMARTS SMILES Arbitrary Target Specification.
SMILES Simplified Molecular Input Entry System.
SMORS Soluble Macromolecular Oxidatively Reactive Species.

SNV Standard Normal Variate.

SPK Synthesized Paraffinic Kerosene.

SPK/A Synthesized Paraffinic Kerosene plus Aromatics.

SVM Support Vector Machine.

SVR Support Vector Regression.

t-SNE t-distributed Stochastic Neighbor Embedding.

TPO Co-processing of pyrolisis ol from used tires.

UHC Unburned Hydrocarbons.

UMAP Uniform Manifold Approximation and Projection.

UMR-PRU Universal Mixing Rule Peng-Robinson UNIFAC.

UV Ultraviolet.

UV-Vis Ultraviolet-Visible.

XGBoost eXtreme Gradient Boosting.

XML Explainable Machine Learning.

Chapter 1

Introduction

1.1 Context of the study

The aviation sector is one of the most important industries in the world. By 2019, it supported around 4.1% of the world's Gross Domestic Product (GDP), transported around 4.5 billion passengers, 0.5% of international shipment by volume and 35% by value, which represented between 35 to 40% by value [1–3]. Furthermore, this sector is the fastest-growing transport mode, and it's expected to grow on average 4.3% annually for the following 20 years. Consequently, aviation is responsible for 12% of the transportation industry's carbon dioxide (CO_2) emissions, while representing 2.1% of all anthropogenic emissions [4]. Besides CO_2, the aviation industry is also responsible for the emission of other gases, such as carbon monoxide (CO), diverse nitrogen oxides (NO_x) and sulfur oxides (SO_x), Unburned Hydrocarbons (UHC), and Particulate Matters (PM) [5]. These gases are air pollutants, also known as Greenhouse Gases (GHG), and are responsible for climate change.

In 2015, representatives from 196 countries gathered to ratify an agreement to limit the impact of human activity on the environment. This treaty, known as the Paris Agreement, has the objective of keeping human-caused global warming under 2 °C, and preferably, under 1.5 °C by the end of the century, compared to pre-industrial levels [6]. In order to set a quantifiable goal for Greenhouse Gases emissions, the concept of "cumulative carbon budget" was defined. The cumulative carbon budget represents the amount of CO_2 that can be emitted to comply with the Paris Agreement goals. For instance, to reach the goals for the +2 °C scenario, aviation emissions should stop by 2075, and for the +1.5 °C, it should reach carbon neutrality by 2050 [1].

Current efforts to reduce the environmental impact of this industry include increasing the efficiency of engines and improving aerodynamics, while future goals aim to develop hydrogen and electric-powered aircrafts. However, the implementation of these technologies

as a solution for CO_2 emissions, faces practical challenges, particularly because 80% of CO_2 emissions originate from long-haul flights exceeding 1500 km, distances that are unfeasible to cover with the current low-energy density batteries. Therefore, advancements in hydrogen fuel storage, lightweight cryogenic tanks, and high-energy density batteries are crucial [1, 2, 7, 8]. As a result, these technologies are not expected to become commercially viable before 2030.

For these reasons, an alternative technology that can be used to reach the sector's climate targets is Sustainable Aviation Fuel (SAF). SAF is a type of aviation fuel produced from renewable or waste-based resources, such as used cooking oil, agricultural residues, municipal waste, or algae, rather than fossil-based sources like crude oil. Eleven conversion pathways are outlined in the ASTM D7566 (Standard Specification for Aviation Turbine Fuel Containing Synthesized Hydrocarbons) [9], and the ASTM D1655 [10] (Standard Specification for Aviation Turbine Fuels).

The chemical composition of SAFs differs from that of conventional fossil fuel. For instance, most SAFs contain minimal or no aromatics, consisting of fewer types of hydrocarbons [11–14]. Thus, SAFs must be blended with fossil jet fuel to ensure compatibility with the existing infrastructure and meet jet fuel specifications. Nevertheless, the blending process might alter the hydrocarbon family ratio of a fuel, potentially affecting its properties [15, 16].

One of the most important parameters for both conventional jet fuels and SAFs is their stability, which is defined as the resistance to degradation processes that can change fuel properties and form undesirable chemical species [17]. The accumulation of said products can lead to system failures at different points. For example, the formation of polar compounds, such as acids, can cause corrosion in engine parts, while the formation of deposits and gums can lead to clogging of engine fuel lines, filters, and injectors[17].

Different approaches have been used to model autoxidation; for example, several detailed mechanisms for different hydrocarbons have been developed [18–20]. However, these approaches cannot consider solvent effects in liquid-phase oxidation. Even though some corrections have been [21, 22] used to incorporate the solvent effects, the number of parameters to be calculated in real fuels with thousands of compounds is computationally unfeasible [23]. An alternative to the current approaches could be using two machine learning-based methods: chemometrics, and cheminformatics. These methods involve algorithms to find relations between a property and the chemical signal or molecular structure of compounds and molecules, respectively. Both approaches have different requirements. For instance, cheminformatic modeling is usually performed from pure compound data, while chemometrics usually uses multivariate data, such as spectroscopic signals.

1.2 Jet fuel

Jet fuel comes from refined kerosene, a petroleum distillate with hydrocarbon chains of C_9-C_{16} and an average boiling point in the range of 150 to 250°C (see table 1.1) [24, 25]. Jet fuel contains approximately 2000 hydrocarbons [11], which are composed by 80% of linear, branched, and cyclic alkanes, also called *n*-paraffins, iso-paraffins and naphthenes, respectively. The remaining 20% is composed of aromatics, such as alkyl-benzenes, and naphthalenes [24]. Heteroatomic species are also present in jet fuel in lower quantities, approximately 1%. Namely nitrogen, oxygen, and sulfur-containing species, such as phenols, indoles, carbazoles, amines, pyridine, anilines, and thiophenes [26–28].

Table 1.1 – Petroleum fractions produced from distillation and their approximate hydrocarbon and boiling ranges[25].

Petroleum fraction	Hydrocarbon range	Boiling range (°C).
Light gases	C_2-C_4	-90 to 1
Gasoline (light and heavy)	C_4-C_{10}	-1 to 200
Naphthas (light and heavy)	C_4-C_{11}	-1 to 205
Jet fuel	C_9-C_{14}	150 to 255
Kerosene	C_{11}-C_{14}	205 to 255
Diesel fuel	C_{11}-C_{16}	205 to 290
Light gas oil	C_{14}-C_{18}	255 to 315
Heavy gas oil	C_{18}-C_{28}	315 to 425
Wax	C_{18}-C_{36}	315 to 500
Lubricating oil	>C_{25}	>400
Vacuum gas oil	C_{28}-C_{55}	425-600
Residuum	>C_{55}	>600

Jet fuel used for civil aviation purposes can be categorized into four main types, each tailored to meet specific performance, safety, and environmental standards [29–31]:

- **Jet A**: Mainly used in the United States, it has a freezing point of -40 °C and doesn't contain a static dissipator additive. However, it is often used interchangeably with Jet A-1.
- **Jet A-1**: The most widely used fuel, it has a flash point minimum of 38 °C and a freezing point of maximum -47 °C. It contains static dissipater additives.
- **Jet B**: A naphtha-kerosene fuel. It is mostly used in cold climates due to its higher flammability and lower freezing point freezing point (-60 °C).

- **TS-1**: Developed to comply with Russian standards, it is considered to be on par with Jet A-1, with the difference of having a lower freezing point (-57 °C).

On the other hand, military aviation fuel is specially formulated to meet the rigorous demands of military operations, including extreme environmental conditions, and enhanced performance requirements. Also, this type of fuel follows the *Single-Fuel Concept* agreement, thus it can be used as a replacement of diesel for land vehicles. Military jet fuel is collectively referred to as Jet Propellant (JP), with the two most important types being [32, 33]:

- **JP-5**: A high flash point kerosene fuel developed in 1952. Primarily used in aircraft carriers.

- **JP-8**: The military equivalent of Jet A-1, primarily used by NATO. It contains several additives, such as corrosion inhibitors, anti-icing agents, metal deactivators and antioxidants. It is used as fuel for both aircraft and ground vehicles.

In an effort to decrease the dependence on petroleum and reduce Greenhouse Gases (GHG) the aviation industry has increasingly turned its attention to renewable synthetic fuels, known as Sustainable Aviation Fuels (SAFs). The adoption of SAF is seen as a pivotal step toward achieving the aviation sector's carbon-neutral growth goals, as outlined by international organizations such as the International Air Transport Association (IATA) and International Civil Aviation Organization (ICAO). One of the defining characteristics of SAF is its versatility in production. It can be synthesized from a wide range of feedstock, including plant-based materials, waste oils, municipal solid waste, and even carbon captured from the atmosphere [34]. Because of this diversity, SAF is not a single fuel but rather a broad category encompassing multiple types of renewable and synthetic aviation fuels.

Depending on the feedstock and conversion processes employed, SAF and other biofuels can be classified into several categories. These categories reflect the variety of technologies and raw materials used in their production [35–37].

- **First generation:** Food-based feedstock, mainly sugars, starches, and oils. Some examples include soybean oil, palm oil and sugarcane. Their main disadvantage is that their production can disrupt the food supply chain while affecting the costs of crops.

- **Second generation:** Fuel obtained from waste and non-food feedstocks, such as lignocellulosic sources. Some feedstocks include cooking oil, beef tallow, sugarcane and forestry residues. Their use can lead to higher GHG emissions than fossil fuel when feedstocks need to be redirected for their production [35].

- **Third generation:** Algae are used to produce oils and sugars, which are later transformed into jet fuel. The main advantages are that algae have no food value and are highly renewable; nevertheless, they haven't been commercially used yet.

- **Fourth generation:** Obtained from non-biological resources, such as CO_2, renewable electricity, water, sunlight, and genetically modified microorganisms. Currently, it is the least mature feedstock.

SAFs are classified as "drop-in" fuels, meaning they can be utilized in existing aircraft fleets without requiring any engine modifications. However, as previously mentioned, SAFs must be blended with conventional jet fuel due to their different chemical composition, ensuring compatibility with aircraft and compliance with the specifications of ASTM D1655. Currently, 11 production pathways have been approved; eight production processes and three pathways for the co-processing of renewable feedstocks in petroleum refineries. ASTM D1655 [10] and ASTM D7566 outline the approved production pathways, feedstocks, and blending limits [9, 10]:

- **Hydroprocessed Synthesized Paraffinic Kerosene (SPK)**: Detailed in ASTM D7566 Annex A1, this fuel is derived from coal, natural gas, or biomass, and can be blended up to 50% with conventional jet fuel. This fuel is produced using the Fischer-Tropsch (FT) process. The FT process converts H_2 and CO, in the presence of an iron or cobalt catalyst, to a wide range of hydrocarbons. Common products include olefins and paraffins in the range of diesel and gasoline [38]. FT is composed by approximately 75% *n*-paraffins and 25% iso-paraffins. This pathway was approved in 2009.

- **Synthesized Paraffinic Kerosene from Hydroprocessed Esters and Fatty Acids (HEFA SPK)**: Described in ASTM D7566 Annex A2, it uses feedstocks like vegetable oils, animal fats, and used cooking oils, with a blending limit of 50%. This pathway consists of a two-step hydrogenation process, (1) hydrotreatment, which removes oxygen by using H_2 in the presence of a catalyst and leads to a paraffinic product, and (2) hydroisomerization, which converts linear paraffins into iso-paraffins in order to improve the cold flow properties of the product [39]. The product's composition depends on the feedstock, but it is mainly made up by iso-paraffins (≈85-90%), *n*-paraffins (≈ 10%) and traces of olefins and naphthenes (less than 1%). It was approved in 2011.

- **Synthesized Iso-paraffins (SIP)**: According to ASTM D7566 Annex A3, this fuel uses sugar-based biomass as feedstock and has a maximum blend ratio of 10%. The production process involves treating the feedstock with genetically engineered microorganisms to produce β-farnesene, a C_{15} tetra-olefin. The unsaturations in β-farnesene are then removed through hydrotreatment, resulting in farnesane as the final product. This fuel is primarily composed of a single compound, with 96–98% farnesane and minor residues of olefins and naphthenes. It was approved in 2014 [40].

- **FT-Synthesized Paraffinic Kerosene plus Aromatics (SPK/A)**: As per ASTM D7566 Annex A4, this allows for blending up to 50%. Aromatics are intentionally added to

FT-SPK for its production. The aromatic content of this fuel ranges from 10 to 20%. Its use was approved in 2015.

- **Alcohol-To-Jet Synthetic Paraffinic Kerosene (ATJ-SPK)**: Listed in ASTM D7566 Annex A5, this fuel is produced from ethanol, isobutanol, or isobutene derived from biomass and supports blends of up to 50%. It is expected that in the future, all C_2 to C_5 alcohols will be permitted for ATJ-SPK production. The production process consists of three main steps: (1) dehydration (applicable only for ethanol and isobutanol), which removes the OH functional group; (2) oligomerization, to increase the chain length of the olefins produced in the first reaction; and (3) hydrogenation, to remove unsaturations. This process is then followed by fractionation [41]. Iso-butanol-based ATJ-SPK was approved in 2016, while ethanol-based ATJ-SPK received approval in 2018. ATJ-SPK primarily consists of iso-paraffins (approximately 99.8%), specifically dodecane isomers and 2,2,4,4,6,8,8-heptamethylnonane.

- **Catalytic Hydrothermolysis Jet (CHJ)**: Detailed in ASTM D7566 Annex A6, CHJ is made from vegetable oils, animal fats, or used cooking oils, with a blend limit of 50%. This fuel is obtained by using the Biofuels Iso-Conversion (BIC) process, which consists of three steps: (1) catalytic hydrothermolysis for cracking and cyclization of triglyceride oils, (2) hydroprocessing for olefin saturation and deoxygenation and (3) fractionation [42]. It was approved in 2020.

- **Synthesized Paraffinic Kerosene from bio-derived Hydroprocessed Hydrocarbons, Esters and Fatty Acids (HC-HEFA SPK)**: Found in ASTM D7566 Annex A7. This fuel must be produced from paraffins obtained from hydrogenation and deoxygenation of bio-derived hydrocarbons, fatty acid esters, free fatty acids. With the recognized bio source being the *Botryococcus braunii* species of algae. It has a blending limit of 10%.

- **Alcohol-To-Jet Synthetic Paraffinic Kerosene with Aromatics (ATJ-SKA)**: Defined in ASTM D7566 Annex A8, it utilizes any single C_2-C_5 alcohol or their combinations. The production involves two subprocesses: a non-aromatic product, obtained using the previously described method for ATJ-SPK, and an aromatic product, produced through dehydration, aromatization, hydrogenation, and fractionation. This fuel can be blended with conventional jet fuel at ratios of up to 50

- **Co-hydroprocessing of esters and fatty acids**: In ASTM D1655 Annex A1, this process involves biomass-based feedstocks like vegetable oils, animal fats, and used cooking oils mixed with petroleum, with a blend limit of 5%.

- **Co-hydroprocessing of Fischer-Tropsch hydrocarbons**: Also in ASTM D1655 Annex A1, it involves Fischer-Tropsch hydrocarbons co-processed with petroleum, with blending restricted to 5%.

- **Co-processing of HEFA**: Listed in ASTM D1655 Annex A1, this fuel uses hydroprocessed esters and fatty acids from biomass, allowing for blending up to 10%.

Blending limits are established to minimize the impact of the different chemical compositions of SAFs on the physico-chemical properties of fuel. Compared to conventional fossil fuels, which contain thousands of compounds, SAFs have simpler compositions. This is illustrated in table 1.2, where HEFA SPK, ATJ-SPK, and SIP are shown to primarily consist of linear, branched, and cyclic paraffins, with minimal aromatic content. Each hydrocarbon family plays a crucial role in determining the overall properties of fuel. For example, aromatics enhance the thermal stability of fuel and contribute to the swelling of rubber seals in aircraft [43].

Table 1.2 – Chemical composition (wt%) of different types of jet fuel [11, 12, 44, 45].

	Jet A-1	**JP-8**	**HEFA SPK**	**ATJ-SPK**	**SIP**
n-paraffins	23.4	26.1	8.5	0.0	0.0
iso-paraffins	27.5	37.5	89.7	99.8	99.5
monocycloparaffins	22.6	19.4	1.7	0.0	0.4
di- and tricycloparaffins	5.8	3.5	0.0	0.0	0.0
alkylbenzenes	14.5	11.0	0.1	0.0	0.1
cycloaromatics	5.1	1.5	0.0	0.0	0.0
alkylnaphthalenes	1.1	1.1	0.0	0.2	0.0

In extreme cases, the blending process can significantly alter fuel properties, requiring blending ratios lower than the specified limits. For instance, blending SPK, HEFA SPK, or SPK/A can influence fuel density or aromatic content, while SIP may affect the viscosity of the refined fuel [9].

Besides the production pathways approved by the ASTM, other conversion processes are currently under evaluation. Some of these candidates include Synthesized Aromatic Kerosene (SAK), Integrated Hydropyrolysis and Hydroconversion (IH2), Single Reactor HEFA (Drop-in Liquid Sustainable Aviation and Automotive Fuel) (DILSAAF), Pyrolysis of non-recyclable plastics (ReOIL), Co-processing of pyrolisis ol from used tires (TPO), Methanol-to-Jet (MTJ), among others [46].

The potential of SAFs to reduce CO_2 emissions largely depends on the production process and feedstock. For example, SPK derived from coal or natural gas can result in an actual increase in emissions compared to fossil fuels, ranging from 10% to 122%. On the other hand, SPK obtained from salicornia and switchgrass can reduce emissions by 93 and 80% respectively, while palm, soy and jatropha oil-derived HEFA SPK limit emissions by 50-60 % during their lifecycle [5]. Thus, the challenge of reducing CO_2 emissions through the use of SAFs involves not only the development of technology but also the availability of feedstock. Furthermore, the

wide adoption of SAFs is severely limited by lengthy approval processes, and high production costs, with 60 to 75% of the cost attributable to the feedstock alone [24].

1.3 Oxidation stability

As previously noted, the distinct composition of SAFs can influence various fuel properties. Of particular interest in this study is their impact on oxidation stability, a critical factor in fuel performance and shelf life. In the context of hydrocarbons, oxidation stability refers to their ability to resist reactions with oxygen [17]. The oxidation process has been extensively studied in the literature and is understood to follow a reaction mechanism initiated by the formation of unstable free radicals through homolytic cleavage of C-H bonds. These free radicals can subsequently undergo various reactions, such as isomerization, β-scission, and addition of O_2.

The study of the oxidation process is of significant interest because the reactions involved can lead to the formation of undesirable chemical species. These include high molecular weight polymers and polar compounds, such as alcohols, carboxylic acids, and water. The accumulation of these species can adversely affect fuel properties, potentially causing engine corrosion and the formation of deposits. Such deposits can clog fuel lines and filters, while the increase in fuel viscosity may accelerate pump wear [47].

Fuel degradation can occur at various stages of its life cycle, including storage, use as a coolant in aircraft systems, or during combustion [48]. This process can occur through one of the following mechanisms [17]:

- Oxidation or autoxidation
- Thermal oxidation
- Hydrolysis
- Microbial contamination

It is worth noting that this classification is subject to debate, since some authors [13] argue that autoxidation is part of thermal oxidation. Thus, thermal oxidation can be further classified into three regimes based on the temperature to which the fuel is exposed:

- **Autoxidation:** < 350 °C
- **Transition:** Between 350 and 450 °C
- **Pyrolitic:** > 450 °C

In this study, we focus on the autoxidation phenomenon, consisting of slow reactions that occur at near-ambient temperatures, where hydrocarbons interact with dissolved oxygen in the bulk of the sample [17]. Given the gradual nature of autoxidation, accelerated oxidation tests have been developed to study this phenomenon in a laboratory setting. Said tests are performed under high temperatures and pressures to accelerate the rate of reactions.

The results of accelerated oxidation tests have provided valuable insights into reactivity trends. When combined with the characterization of oxidation products, these findings have formed the foundation for the development of reaction mechanisms [49, 50]. In the next section, we will discuss the various types of reaction mechanisms studied in the field of chemical kinetics.

1.4 Kinetic modeling

We can consider that a model is a type of "box" that transforms inputs into outputs through relations [51]. Modeling approaches can be categorized in three types: white, black and gray-box. White-box modeling, also known as first-principles or analytical modeling, relies on established relationships and knowledge of internal mechanisms to construct a model [52]. In contrast, black-box modeling, or data-driven modeling, directly derives relationships between inputs and outputs from experimental or historical data. This approach is particularly useful when the system's behavior is not fully understood, or when dealing with complex systems that are challenging to analyze using white-box methods [53]. Gray-box modeling bridges these two approaches by requiring partial knowledge of the system. In this hybrid method, well-understood components are analyzed using analytical techniques, while data-driven modeling is applied to less understood aspects [54].

Kinetic modeling, an analytical or white-box approach, is a key method for studying the dynamics and mechanisms of chemical reactions often used in conjunction with experimental techniques to provide a comprehensive understanding. Kinetic modeling predicts the rate of chemical reactions, as well as changes in reactants and products concentrations over time. In the context of autoxidation, empirical information can be obtained from the results of oxidation tests.

Chemical kinetics studies how experimental conditions affect the rate of a chemical reaction, uncovering details about its mechanism and transition states. For this purpose, mathematical models that describe the relevant factors of a chemical reaction are used. Kinetic modeling requires identifying a set of reactants and products, along with the chemical reactions that connect them [55]. However, accounting for all the chemical species and reactions involved in a process is a complex task. To manage this, model reduction techniques such as quasi-equilibrium states, limiting steps, and subsystems, are employed to simplify the model's complexity [56]. Another approach for model reduction involves grouping chemical reagents into classes. We now proceed to discuss the various types of kinetic modeling based on different levels of reagent grouping.

1.4.1 Global modeling

Global mechanisms are used to represent complex systems as a single chemical species. In the context of oxidation, a given hydrocarbon or fuel can be denoted as RH, while the oxidation mechanism is usually represented by its main stages (initiation, propagation and termination), and oxidation products are grouped by species, such as alkyl and peroxyl radicals, hydroperoxides, etc. For instance, the H-abstraction for a generic jet-fuel could be written as:

$$RH + O_2 \longrightarrow R^{\bullet} + HO_2^{\bullet} \tag{1.1}$$

Reaction kinetics in this approach are considered to follow the Arrhenius equation [57, 58]:

$$k = A \exp^{-\frac{E_a}{RT}}, \tag{1.2}$$

where:

- k is the kinetic constant,
- A, the pre-exponential factor,
- E_a, the activation energy,
- R, the gas constant, and
- T, the temperature in Kelvin.

In global modeling, these kinetic parameters are calculated by fitting the Arrhenius equation to experimental data, such as reagent or product concentrations [58, 59].

The main advantage of this approach is that it allows to study complex matrices, such as conventional and alternative jet fuels [60], diesel and biodiesel [61, 62], vegetable oil [59], gasoline [58], among others. On the other hand, this type of modeling is limited by its simplicity, since it doesn't consider the influence of alternative reactions, and the effects of other chemical compounds, such as antioxidants and heteroatomic species [63].

1.4.2 Semi-detailed modeling

Semi-detailed modeling builds upon global models by extending the mechanism to account for interactions with other chemical species. For this purpose, compounds are grouped into classes called "lumped species", such as as metals, anti-oxidants, oxygenated species, sulfur-containing molecules and nitrogenated compounds [64]. Furthermore, this approach includes additional mechanism pathways, such as secondary oxidation reactions, e.g. hydroperoxide decomposition into alcohols and other oxygenated species, and chain reaction termination, by recombination of $R^{\bullet}$ and $ROO^{\bullet}$ radicals. These chemical reactions are detailed in Chapter 2, in equations (2.5) to (2.7).

Several semi-detailed mechanisms have been proposed for various hydrocarbons. For instance, the oxidation of indene and tetralin in the presence of free radical scavengers was studied by Carlsson and Robb [65]. Garcia-Ochoa et al. [66] modeled the thermal oxidation of *n*-octane, incorporating hydroperoxides, ketones, alcohols, and acids into the mechanism. Similarly, Blaine and Savage [67] developed a 12-reaction mechanism for *n*-hexadecane autoxidation, which included hydroperoxides and secondary oxidation products such as ketones, alcohols, acids, and esters. For aromatic compounds, Hermans et al. [68] and Hoorn et al. [69] proposed mechanisms for the autoxidation of toluene. Said mechanisms included the formation of benzaldehyde, benzyl alcohol, and benzoic acid.

Some of the first mechanisms for jet fuel autoxidation and deposition were developed by Zabarnick [70]. In these studies, the authors investigated the impact of naturally occurring antioxidant species on the thermal and oxidation stability of jet fuel [63, 70]. Nevertheless, certain deficiencies in the modeling approach became evident, such as the poor fit for two-parameter Arrhenius global oxidation reactions and the system's non-Arrhenius behavior [71].

In more recent work, the study of Kuprowicz et al. [64] proposed a 18-reaction mechanism for the prediction of autoxidation and deposition of jet fuels. In their study, the authors considered chemical species, such as hydrocarbons (R), dissolved oxygen (O_2), peroxyl radical inhibitors or antioxidants (AH), reactive sulfur species (SH) that act as hydroperoxide decomposers, and hydroperoxides (ROOH). This mechanism was modified by Sander et al. [48], who updated the activation energy values for the unimolecular decomposition reaction of peroxy radicals, $RO_2^{\bullet}$, and removed some reactions, among other changes.

While semi-detailed mechanisms broaden the scope of global models by incorporating secondary chemical reactions, the grouping of compounds into classes or lumped species may overlook the effects of minor components. Moreover, these mechanisms are often influenced by the experimental conditions under which the data were obtained [67].

1.4.3 Detailed modeling

A detailed mechanism is composed of several elementary steps or reactions, which can be defined as the transition from a set of reactant molecular structures to a set of product molecular structures. These reactions are only dependent on chemical structure and sometimes, pressure and temperature [72]. Normally, the obtained mechanisms involve hundreds of chemical species and thousands of reactions, grouped into categories for ease of interpretation. Most of the efforts on this type of modeling has been focused on the study of gas-phase oxidation. For instance, detailed mechanisms have been developed for heavy *n*-alkanes, ranging from C_8 to C_{16} [73], decalin [19, 74], *n*-hexadecane and iso-cetane [75], and toluene [76]. For fuels such as diesel [18, 20] and jet fuel [20], modeling has been performed by representing them as surrogate

mixtures. More recently, Dong et al. [77] developed a detailed kinetic model for surrogate fuels, which includes linear C_8-C_{12} alkanes, PAHs (Polycyclic Aromatic Hydrocarbons), NO_x pollutants, among other chemical species.

In order to account for all the possible reactions, a reaction mechanism generator is used. Some examples include NetGen [78], MAMOX [79], EXGAS [80], Genesys [81] and RMG [82]. These software generate feasible reaction mechanisms by estimating thermodynamic and kinetic parameters on gas phase with theoretical methods, such as DFT calculations. Normally, a perfect gas behavior of the involved chemical species is assumed, indicating that no interaction between molecules is considered and solvent interactions do not influence thermodynamic parameters [23].

However, the assumption of non-interaction in the ideal gas model does not apply to liquids. Thus, gas-phase kinetic data need to be corrected for liquid-phase systems. Some efforts have been put into achieving this. For instance, the development of the continuum solvation models like Polarizable Continuum Model (PCM) [83], COnductor-like Screening Calculation MOdel (COSMO)[84]. Furthermore, Jalan et al. [85] developed a framework that relied on calculating Gibbs free energy of solvation ($\Delta_{solv}G(T)$) corrections on gas-phase data, by using a Linear Solvation Energy Relationship (LSER) combined with a group additivity method. Le et al. [23] proposed an alternative approach for the calculation of Δ_{solv}. Their work consisted on using Equation of State (Equation of State) Universal Mixing Rule Peng-Robinson UNIFAC (UMR-PRU) [86] to calculate $\Delta_{solv}G(T)$ as a function of temperature for the chemical species in the mechanism.

Nevertheless, the main limitation of these frameworks is the high number of elementary reactions, and thermodynamic and kinetic parameters that need to be calculated for a single chemical species. It is evident that for a complex mixture containing thousands of compounds, the calculation of all the parameters becomes computationally unfeasible [23]. An attempt to circumvent this limitation requires representing real fuel as a surrogate of 2 to 6 compounds, that are representative of the fuel chemical composition and properties, such as ignition delay time, laminar flame speed, engine combustion, cetane number, distillation curve, etc [20]. Regardless, this approach may be an oversimplification that doesn't represent the chemical composition variety in the fuel. Due to said limitations, we decided to explore data-driven modeling approaches, that may allow to estimate the oxidation stability of complex fuels without the inherent drawbacks of detailed-modeling; computationally unfeasible calculations and solvent-dependent corrections.

1.5 Aims and objectives

The aim of the present work is to use data-driven approaches, such as chemometrics and cheminformatics, to predict and understand the oxidation stability under the autoxidation regime of hydrocarbons relevant to conventional and alternative jet fuels.

The objectives of the present work are:

- Generate a database from accelerated oxidation measurements of pure hydrocarbons.
- Perform spectroscopic measurements of fresh and oxidized hydrocarbons.
- Develop a cheminformatic model based on molecular descriptors to predict the oxidation stability.
- Develop a chemometric model based on the spectroscopic signals to predict the oxidation stability.
- Identify relevant features, chemical descriptors and spectroscopic regions, related to oxidation stability.

This manuscript is organized as follows: Chapter 1, the current chapter, provides a general overview of the topic and highlights the gap in the literature that this work seeks to address. Chapter 2 discusses oxidation stability, the experimental methods employed in its study, and presents the results of the accelerated oxidation tests we conducted, which led to the identification of new reactivity trends. In Chapter 3, we present a model based on molecular descriptors, developed using cheminformatics tools, and discuss the key molecular features identified by the model. Chapter 4 focuses on the Near-Infrared spectra of our samples and the predictive model derived from this spectral data. Finally, in Chapter 5, we summarize our findings and discuss future perspectives.

Chapter 2

Liquid-phase oxidation of hydrocarbons

2.1 Introduction

As previously mentioned, oxidation stability refers to the resistance of hydrocarbons to react with oxygen [17]. This property has been shown to be closely linked to the molecular structure of hydrocarbons, with numerous studies focusing on uncovering relationships between oxidation stability and specific molecular features. For example, Stephens and Roduta [49] and Larsen et al. [50] developed experimental methods for studying the oxygen consumption of several paraffinic and aromatic compounds subjected to thermal stress. These works support that the stability of paraffins decreases with respect to their chain length and the effect of the bonding patterns in alkyl carbons attached to aromatic rings. Nevertheless, these studies required the use of in-house developed methods, posing reproducibility issues and making difficult the comparison of literature data. Thus, to circumvent these limitations, standardized methods, such as the Rapid Small Scale Oxidation Test (RSSOT) [87, 88], also known as the PetroOxy or RapidOxy test, have been developed.

Similarly, several authors have employed the RapidOxy instrument to analyze a variety of hydrocarbons, and found some correlations between the oxidation stability and molecular features. For instance, Skolniak et al. [89] found that long paraffinic chains, unsaturations, and aliphatic rings were related to low oxidation stability. On the other hand, the authors observed that unsubstituted aromatic rings and branched paraffins were correlated with high stability. Chatelain et al. [90, 91] corroborated the conclusions regarding the stability of linear paraffins' stability, but discovered that the stability of branched paraffins depends on the bonding pattern or connectivity of the carbon atoms available in the molecule. Thus, the C centers exhibit the following stability order: quaternary > primary > secondary > tertiary.

Ben Amara et al. [16] performed studies on the stability of aromatic and di-aromatic compounds and their saturated counterparts. The authors found that di-aromatics are more

stable than mono-aromatics, while the latter are more stable than cyclic paraffins, concluding, thus, that the length of the paraffinic chains attached to aromatic rings have a negative impact on their stability. However, they also concluded that the number of substituents of the aromatic ring does not influence the reactivity of the molecules, which is in disagreement with the findings of Stephens [92], who observed that 1,2,4,5-tetramethylbenzene (durene) is more easily oxidized than mesitylene, *m*-xylene, and toluene. We believe that these seemingly contradictory conclusions regarding the structure-property relationships may stem from the analysis of a limited number of hydrocarbons. Furthermore, there are a many relationships that remain unexplored, such as the effect of substituent position in aromatic compounds or the substitution order of alkenes.

This work has the objective of extending the structure-property relationships found in the literature by performing accelerated oxidation tests. To ensure repeatability and obtaining comparable results, we used the PetroOxy/RapidOxy reference method. The analyzed samples span different carbon number ranges and hydrocarbon families, such as paraffins, olefins, naphthenes, aromatics and di-aromatics. Furthermore, we compared our results with the data available and proceeded to identify new critical molecular features related to oxidation stability.

2.2 Autoxidation and oxidation stability

2.2.1 Autoxidation mechanism

At low temperatures, oxygen dissolved in the fuel reacts to form hydroperoxides, unstable species that further react with the fuel bulk, resulting in the formation of deposits and polar species. This process is known as autoxidation [17, 93, 94]. The autoxidation process can be classified into two stages: primary and secondary oxidation. It is widely accepted that primary oxidation proceeds via a radical chain reaction [13, 17, 57, 63, 70, 95–98], following the mechanism outlined below:

- **Step 1:** Initiation
  - $I \longrightarrow I^{\bullet}$ (2.1)
  - $I^{\bullet} + RH \longrightarrow R^{\bullet} + IH$ (2.2)
- **Step 2:** Propagation
  - $R^{\bullet} + O_2 \longrightarrow RO_2^{\bullet}$ (2.3)
  - $RO_2^{\bullet} + RH \longrightarrow ROOH + R^{\bullet}$ (2.4)
- **Step 3:** Termination
  - $R^{\bullet} + R^{\bullet} \longrightarrow R{-}R$ (2.5)

- $R^{\bullet} + RO_2^{\bullet} \longrightarrow ROOR$ (2.6)
- $ROO^{\bullet} + ROO^{\bullet} \longrightarrow$ Molecular products (2.7)

The initiation step consists of the abstraction of a hydrogen atom from a hydrocarbon molecule (RH) in order to generate a free radical. This process can be facilitated by the presence of an initiator (I), which produces initiator radicals ($I^{\bullet}$) through one of the following mechanisms [17, 57]:

- Thermal dissociation of peroxides and hydroperoxides (ROOR and ROOH):
  - $R^1O{-}OR^2 \longrightarrow R^1O^{\bullet} + R^2O^{\bullet}$ (2.8)
  - $RO{-}OH \longrightarrow RO^{\bullet} + OH^{\bullet}$ (2.9)
- Metal catalyzed decomposition of peroxides and hydroperoxides:
  - $ROOH + M^{2+} \longrightarrow RO^{\bullet} + OH^{\bullet} + M^{3+}$ (2.10)
  - $M^{3+} + ROOH \longrightarrow ROO^{\bullet} + H^{\bullet} + M^{2+}$ (2.11)
- Photo-oxidation: catalysis due to fuel exposure to light.

In general, it is considered that metallic ions are one of the main drivers of the initiation reactor [17]. These impurities are introduced into the fuel during the manufacturing process and throughout the supply chain, as the presence of polar species, such as naphthenates and naphthenic acid, promotes their dissolution in the fuel [96]. Studies have shown that even at trace concentrations (ppm to ppb), metals such as copper (Cu) and iron (Fe) can significantly impact the oxidation stability of fuels [99, 100]. For instance, Sarin et al. [101] studied the impact of five different metal traces (Fe, Ni, Mn, Co and Cu) on the stability of Jatropha biodiesel. Their results showed that among the contaminants, Cu and Co accelerated the oxidation of the studied samples by a factor of 8, at concentrations as low as 1.5 ppm. However, the effect of metals can be mitigated by the addition of metal deactivators, which can reduce the catalytic properties of metals by chelating the dissolved ions, passivating metal surfaces and through bulk phase reactions [102].

During the second step, propagation, one of the radicals ($R^{\bullet}$) reacts with molecular oxygen to form a peroxyl radical ($ROO^{\bullet}$). The resulting product is unstable and reacts with an unoxidized molecule, producing a carboxylic acid, and another free radical [17]. The propagation of chain radicals is possible since the previously described reactions are involved in a cyclic sequence, continuously generating radicals in the process [57]. Termination is the last step of primary autoxidation. This stage occurs when the concentration of free radicals is sufficiently high for these compounds to react with each other. For instance, at low temperatures, peroxyl radicals may produce peroxyl linked molecules, by following the reaction:

$$\mathrm{ROO^{\bullet} + ROO^{\bullet} \longrightarrow R{-}OO{-}R + O_2}. \tag{2.12}$$

While termination marks the end of primary oxidation, the oxidation process continues. During secondary oxidation, hydroperoxides decompose, resulting in the formation of soluble and stable molecular products, such as ketones, aldehydes, acids, esters, lactones, furanones, epoxides and polymeric species [13, 17, 57]. However, some studies have shown that the formation of said products may occur much earlier in the chain-reaction. In liquid-phase, the formation of solvent cages promote the reaction between peroxyl radicals and labile αH atoms of hydroperoxides, resulting in the production of ketones (Q=O) and alcohols [68, 103].

$$\mathrm{ROO^{\bullet} + ROOH \longrightarrow ROOH + R_{-\alpha H}{}^{\bullet}OOH \longrightarrow ROOH + Q{=}O + {}^{\bullet}OH}. \tag{2.13}$$

Further oxidation results in the formation of Soluble Macromolecular Oxidatively Reactive Species (SMORS), which originate from the reaction combination of the previously mentioned soluble products, hydroperoxides and heteroatomic compounds present in the bulk of the fuel. SMORS are compounds with high polarity that can undergo further polymerization, increasing their size until becoming insoluble and leading to the formation of deposits [13, 104].

A common way of slowing the oxidation process is the addition of antioxidants. There are two types of antioxidants: hydroperoxide decomposers and chain breakers. The first, as their name suggests, decompose hydroperoxides into more stable compounds, such as alcohols, while the antioxidant turns into an innocuous oxidized form [17]. On the other hand, chain breakers disrupt the autoxidation chain reaction mechanism by intercepting hydroperoxides at a higher rate than the substrate. The most common antioxidants are phenolic species (ArOH) , with Butylated Hydroxytoluene (BHT) being the most commonly used compound [105]. ArOH react with free radicals through the following reaction:

$$\mathrm{ROO^{\bullet} + ArOH \longrightarrow ROOH + ArO^{\bullet}}. \tag{2.14}$$

In the case of BHT, its efficacy as an antioxidant is caused by a series of factors. Firstly, the low O-H Bond-Dissociation Energy (BDE), which permits a fast reaction with free radicals. Secondly, the stabilization of the resulting phenoxyl radical $\mathrm{ArO^{\bullet}}$ through inductive and hyperconjugative effects. And lastly, steric hindrance caused by the two *tert*-butyl groups, which prevent further reactions of the phenoxyl radical, and consequently, the propagation of the chain reaction [106]. However, the use of antioxidants can have negative effects, such as increasing deposit formation [26]. A mechanism for a generic antioxidant A, was proposed by Heneghan and Zabarnick [63], with equation (2.17) being the major production pathway for solids.

$$A^{\bullet} + O_2 \longrightarrow AO_2^{\bullet} \tag{2.15}$$

$$AO_2^{\bullet} + AH \longrightarrow AO_2H^{\bullet} + A^{\bullet} \tag{2.16}$$

$$AO_2^{\bullet} + AO_2^{\bullet} \longrightarrow \text{products.} \tag{2.17}$$

2.2.2 Molecular features related to oxidation stability

Many studies have been performed to determine the stability of different hydrocarbon families present in fuels, in order to identify trends among them. As a general rule, it has been found that the oxidation stability follows the trend: naphthalenes (di-aromatics) > mono-aromatics > *n*-paraffins > naphthenes (cycloalkanes) > iso-paraffins (except containing quaternary carbons) [13], however, there are several cases where this trend doesn't hold.

In the case of linear paraffins, it has been shown that the oxidation stability is inversely proportional to the chain length [50, 89, 90]. Specifically, a non-linear dependence between the stability and the number of carbons has been observed in the C_6 - C_{16} range (see figure 2.1**a**) [91]. This non-linear behavior could be explained by the increased concentration of reactive species caused by the fragmentation of parent molecules [107]. Skolniak et al. [89] found that the oxidation products of *n*-heptane were alcohols and ketones with carbon number of 7, however *n*-hexadecane's oxidation products included compounds with lower carbon numbers, ranging from C_7 to C_{12}.

Regarding the stability of branched paraffins, Mielczarek [108] found that iso-alkanes are more reactive than their linear counterparts. These results were supported by the work of Skolniak et al. [89]. However, the authors also found that isooctane (2,2,4-trimethylpentane) was 10 times more stable than *n*-octane, concluding that this increase in reactivity was caused by steric effects impeding the access of O_2 to the molecule. In another related study performed by Chatelain et al. [91], the authors studied the stability of C_8 isomers: *n*-octane, 2-methylheptane, 2,5-dimethylhexane and isooctane. It was observed that the stability of the compounds decreased with the number of ramifications, however isooctane presented an enhanced stability, similar to that of *n*-octane. The authors explained that the stability of isooctane is caused by the absence of H atoms at the branching sites, rather than steric effects, leading them to conclude that the reactivity of the substituted site depends on the C-H bond dissociation enthalpy and the stability of the resulting alkyl radicals. Thus, the following reactivity trend has been suggested: tertiary carbon > secondary carbon > primary carbon (see figure 2.1**b**) [13, 91].

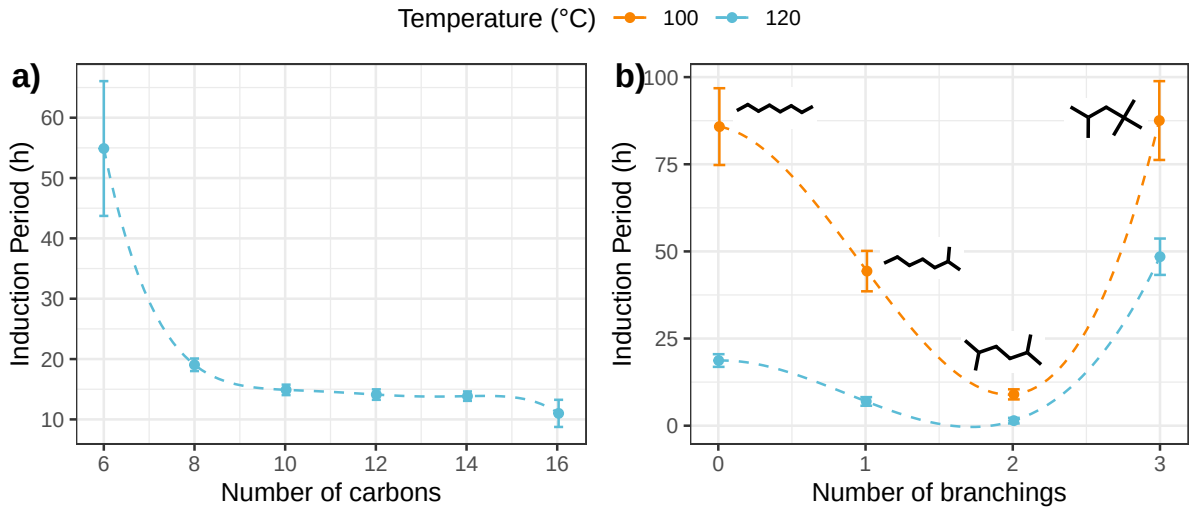


FIGURE 2.1 – **a)** Oxidation stability of *n*-paraffins as a function of the chain length, adapted from Chatelain et al. [90]. **b)** Oxidation stability of iso-paraffins as a function of the number of branching, adapted from Chatelain et al. [91].

Conversely, naphthenes (cyclo-alkanes) tend to be more reactive than their linear counterparts. For instance, it has been found that cyclohexane is twice more reactive than *n*-hexane [89]. Also, the reactivity of these compounds is linked to the number of rings in their structure; in general, their stability is as follows: mono-cyclic > di-cyclic > tri-cyclic [50]. Furthermore, the addition of a paraffin side chain to the ring increases the compound reactivity, for example, octadecyldecalin consumed 1.6 times more oxygen than decalin, while 9,10-diisobutylperhydroanthracene consumed 5 times more oxygen than perhydroanthracene at constant time [50] (see figure 2.2).

Aromatics have been found to be more stable than their linear and cyclic counterparts. At 140 °C, benzene is 1.7 times more stable than hexane and 3.5 times more stable than cyclohexane [89]. However, the stability of aromatics is severely affected by the substitutions on the aromatic ring. For example, it has been found that *n*-propylbenzene is twice as reactive as toluene [16], and just as with linear paraffins, long chains are related to high reactivity [50]. This behavior can be explained by the activation of α hydrogen atoms in the side chain, which are more prone to oxygen addition. Proof for this hypothesis includes the frequent characterization of benzaldehyde in oxidation products, suggesting that the phenyl ring is not attacked [50]. Furthermore, this hypothesis is backed up by experiments on several alkylbenzenes substituted with different carbon types. Larsen et al. [50] found that the stability of *n*-pentylbenzene increased with substitution on the carbon adjacent to the phenyl group. Furthermore, they found that *tert*-pentylbenzene, a compound with a quaternary carbon and no α hydrogen was 30 times more stable than *n*-pentylbenzene, while *sec*-butylbenzene was only 3 times more stable (see figure 2.3) [50]. However, contrary to logic, the number of substituents in the phenyl group doesn't significantly impact the oxidation stability [16, 50].

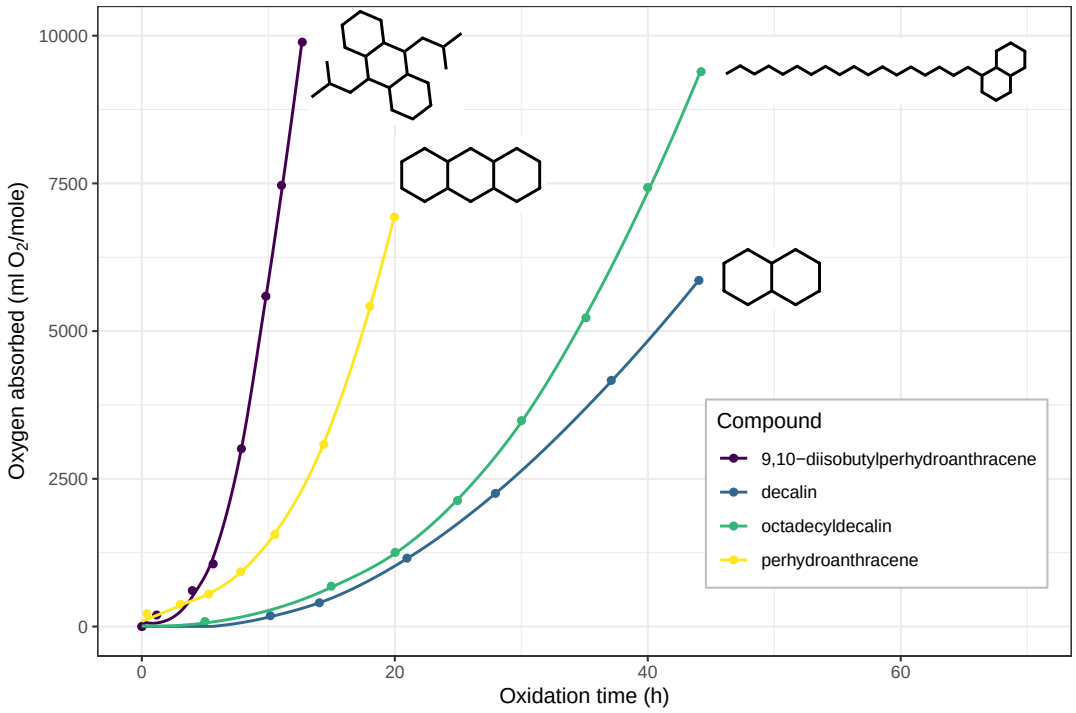


FIGURE 2.2 – Oxidation of naphthenes at 110 °C. Adapted from Larsen et al. [50].

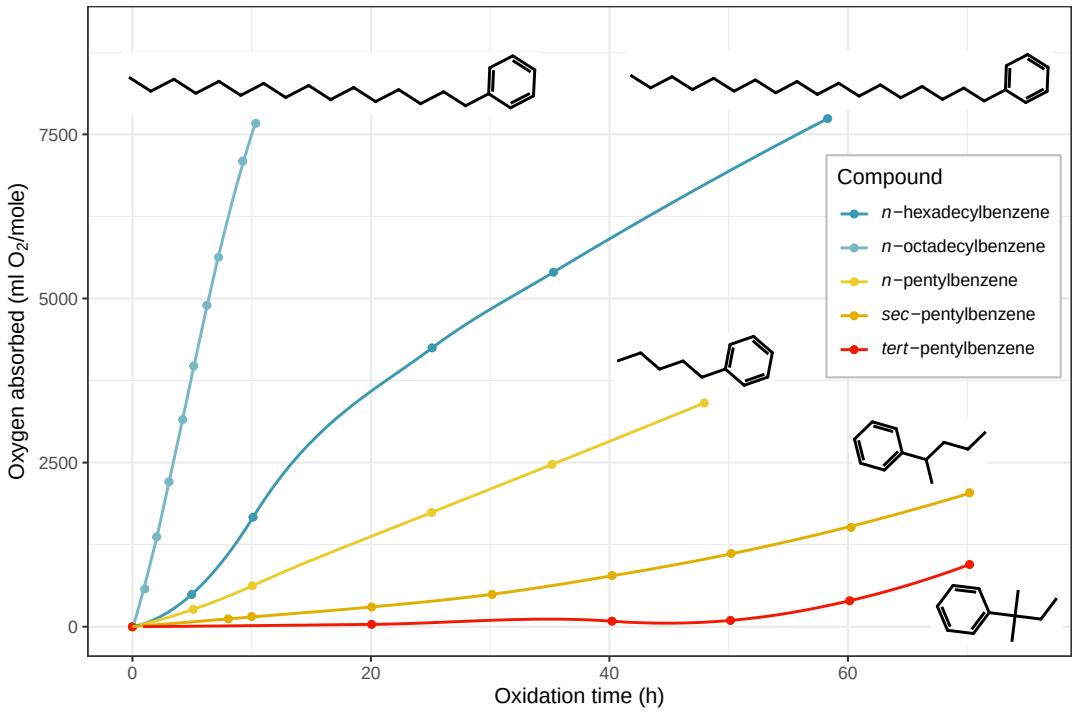


FIGURE 2.3 – Oxidation of alkyl-benzenes at 110 °C. Adapted from Larsen et al. [50].

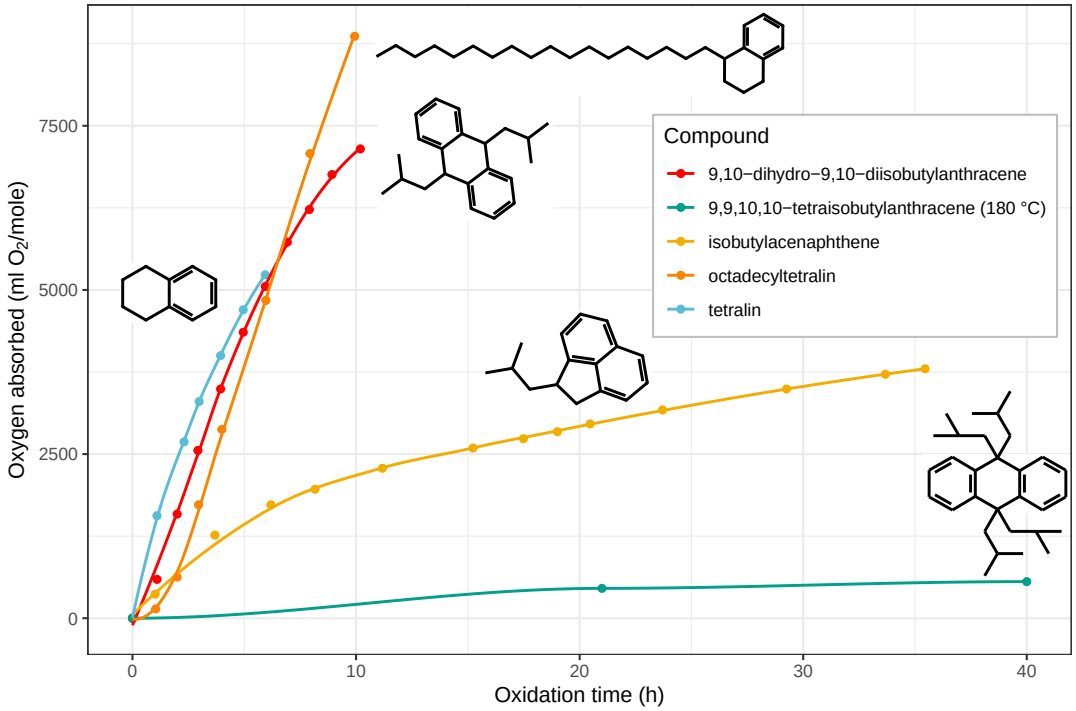


FIGURE 2.4 – Oxidation of naphtheno-aromatics at 110 °C, 9,9,10,10-tetraisobutylanthracene was oxidized at 180 °C. Adapted from Larsen et al. [50].

Naphtheno-aromatics' reactivity is related to that of alkyl-aromatics and naphthenes. These compounds present very high reactivity due to the double substitution on the phenyl group and the presence of four activated α hydrogens [109]. For example, tetralin is one of the most reactive species studied in the bibliography [16, 50]. However, when the carbon atoms adjacent to the aromatic group are substituted, the stability of the compound increases, for example, 9,9,10,10-tetraisobutylanthracene is more stable than 9,10-diisobutylanthracene (see figure 2.4), further validating the hypothesis that activated α hydrogens are responsible of reactivity [50].

Diaromatics or naphthalenes are known for their high stability. For instance 1-methylnaphthalene is twice as stable as toluene [16]. This behavior is caused by naphthalene's auto-retardating and self-inhibiting effects. The auto-retardating effect is caused by the formation of methylnaphthoquinones, which trap alkyl radicals, $R^{\bullet}$ more effectively than peroxyl radicals $ROO^{\bullet}$. On the other hand, the self-inhibiting effect could be caused by first order termination due to methylnaphthoxyl radicals [110]. It is also hypothesized that naphthols are formed during oxidation, thus acting as antioxidants [109]. Like aromatics, naphthalenes are more stable than their substituted counterparts. Reactivity increases with the side chain length, and quaternary carbon substituents (*tert*-butyl, *tert*-pentyl and *tert*-octyl naphthalenes) present higher stabilities than linear chains. Additionally, the position of the substituent influences reactivity, with substitutions at the α position being more reactive than those at the β position (see figure 2.5) [50].

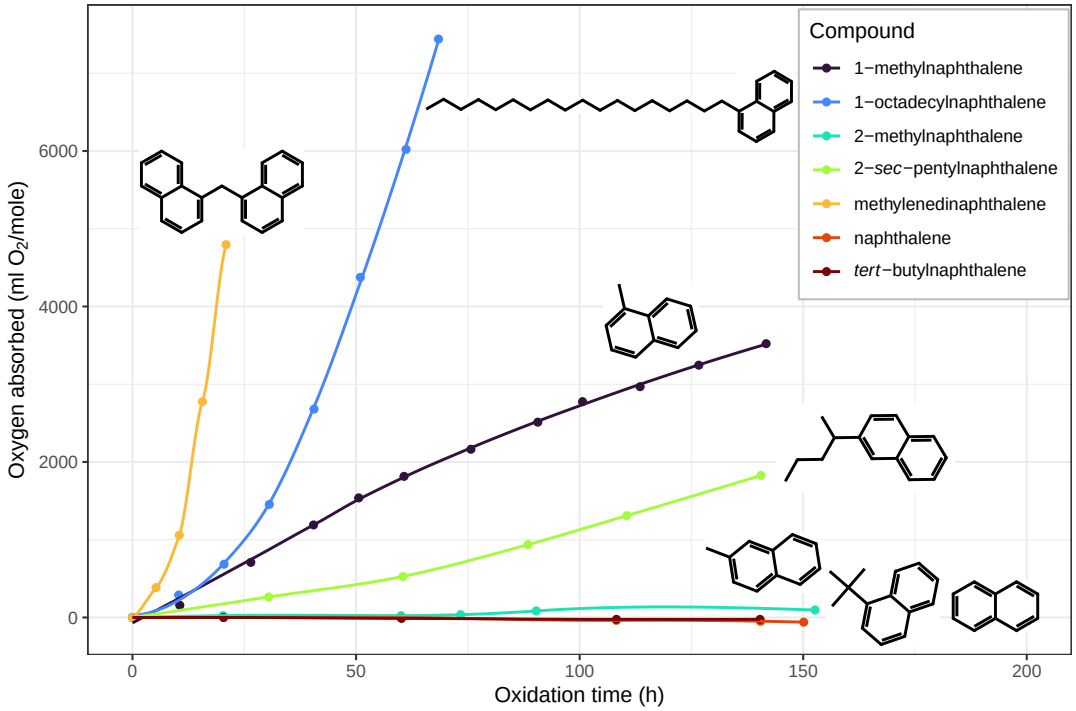


FIGURE 2.5 – Oxidation of naphthalene and its derivates at 150 °C. Adapted from Larsen et al. [50].

Olefins or alkenes are known for their low oxidation stability. For instance, Skolniak et al. [89] observed that 1-hexene is 33 times more reactive than *n*-hexane. The enhanced reactivity of olefinic compounds has been related to the hydrogen atoms bonded to allylic carbons, which are more easily abstracted by oxygen than those bonded to non-allylic carbons. This is due to the resonance stability provided by the adjacent pi electron system. Furthermore, it has been observed that methylene groups that are allylic with respect to two double bonds, or bis-allylic present a higher reactivity that allylic sites [111–113].

2.3 Experimental characterization of oxidation stability

Oxidation stability is not a property that can be directly measured, nevertheless, several chemical indices can be related to it, most of them describe structural aspects of hydrocarbons present in fuels, while others quantify impurities related to instability. While the chemical indices presented in the following sections describe part of the chemical instability of fuels, no single index can account for all the reactivity.

2.3.1 Fuel quality parameters

2.3.1.1 Peroxide value

As discussed in section 2.2.1, hydroperoxides are the main product of primary oxidation. Therefore, quantifying these species is a key method for assessing the oxidation level of a sample. The chemical index associated with this property is the peroxide value, expressed in milliequivalents of peroxide per kilogram of sample. This value is determined using the procedure outlined in the Standard Test Method for Hydroperoxide Number of Aviation Turbine Fuels, Gasoline, and Diesel Fuels (ASTM D3703) [114].

The ASTM D3703 standard uses iodometry to determine the peroxide value. In this technique, hydroperoxides (ROOH) react with iodide (I^-) in an acidic medium to produce iodine (I_2), according to equation (2.18), which is then titrated with an aqueous thiosulfate solution (equation (2.19)) [115].

$$ROOH + 2\,I^- + 2\,H^+ \longrightarrow ROH + H_2O + I_2 \tag{2.18}$$

$$I_2 + 2\,S_2O_3^{2-} \longrightarrow S_4O_6^{2-} + 2\,I^- \tag{2.19}$$

However, the method detailed in the ASTM D3703 suffers from repeatability issues due to the miscibility of the titrant solution in the organic sample. For this reason, Roohi and Rajabi [115] proposed a water-free titration technique that uses triphenylphosphine (TPP) dissolved in *n*-decane as titration solution. TPP reacts free iodine to form an uncolored complex $[TPPI]^+[I]^-$ (see equation (2.20)). Thus, the equivalence point is obtained from colorimetric titration.

$$I_2 + TPP \longrightarrow [TPPI]^+[I]^- \tag{2.20}$$

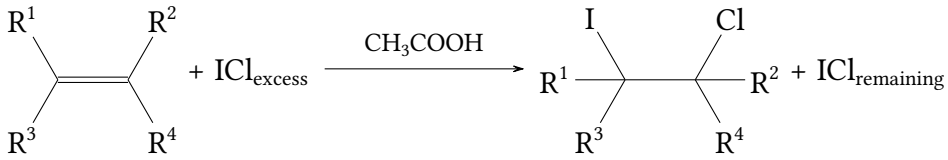
Recently Benrabah et al. [116] adapted the protocol reported by Roohi and Rajabi [115], implementing the detection of the equivalence point through potentiometric titration. This modification resulted in a decrease of the limit of detection by two orders of magnitude with respect to the original implementation. Furthermore, Benrabah et al. [116] developed a HPLC system coupled to a post-column reactor and a Ultraviolet-Visible (UV-Vis) detector, which permits the detection of hydroperoxide concentrations as low as 10^{-10} mol/L.

Despite the improvements for the detection of peroxide content, this chemical index remains a limited measure of oxidation. This is caused by the instability of peroxides, which spontaneously decompose into stable molecules [17, 117].

2.3.1.2 Iodine value

Quantifying the degree of unsaturation is crucial, as it can significantly affect oxidation stability [89]. The Iodine Value (IV) measures the level of unsaturation in oils, fats, and waxes. It is expressed as the mass of iodine, in grams, that reacts with 100 grams of the sample. The official method of determination is known as Wij's method, a titration method that uses iodine monochloride (ICl) or iodine monobromide (IBr) in acetic acid (CH_3COOH), known as Wij's solution.

The ICl or IBr present in Wij's solution reacts in excess with the double bonds of the sample, then the remaining ICl or IBr reacts with KI, resulting in the formation of iodine, which is then titrated with a sodium thiosulfate solution (see equations (2.21) to (2.23)) [118].



(2.21)

$$KI + ICl_{remaining} \longrightarrow I_2 + KCl \tag{2.22}$$

$$I_2 + 2\,S_2O_3^{2-} \longrightarrow 2\,I^- + S_4O_6 \tag{2.23}$$

The primary limitations of this method are that iodine does not react stoichiometrically with conjugated double bonds, and it fails to provide information about the location of the unsaturations, which can significantly influence oxidation stability [17].

2.3.1.3 Total contamination

The gravimetric method described in the European Standard (EN) 12662 is used for determining the amount of insoluble contaminants biodiesel. During the test, a weighted amount of sample is filtered under vacuum through a filter with specified porosity. Then, the filter is washed, dried and weighted, finally, the contamination is expressed as mg of contaminant per kg of sample [17].

2.3.1.4 Acid Number

Acid number indicates the quantity of fatty and mineral acids in a fuel sample. High fuel acidity is related to oxidation products, corrosion and engine deposits. It is determined according to the

volumetric test method EN 14104, by using dilute KOH in ethanol solution and a potentiometer, it is expressed in mg of KOH needed to neutralize 1 gram of sample [17].

2.3.1.5 Viscosity and density

Kinematic viscosity is a parameter that changes throughout the fuel oxidation process. It is determined according to the EN ISO 3104 standard.

During secondary oxidation, molecules can form dimers, trimers, and oligomers, leading to an increase in the fuel's viscosity and density [17]. However, some studies performed on oils have shown that used samples tend to have lower viscosities than fresh oils. Thus, it has been suggested that the cleavage of C-C bonds during the oxidation process results in the formation of low molecular mass compounds, which result in a decrease on viscosity [119].

Density is determined following the procedure outlined in the standard test method EN ISO 3675, using a temperature-controlled bath [17]. Engeländer et al. [120] studied the evolution of biofuel density over 18 months at 8 °C and 40 °C. They found that, on average, samples stored at 8 °C showed a 0.2% increase in density, while those stored at 40 °C exhibited an increase of up to 0.9%.

2.3.2 Accelerated oxidation tests

Autoxidation is a long process that takes months to years under ambient conditions. To study this phenomenon in a laboratory setting, accelerated experiments are conducted under controlled temperature and pressure conditions. Early studies employed rudimentary setups for these experiments. For example, Stephens et al. [49, 92, 121–123]. performed a series of investigations on the reactivity of aromatic compounds. For this purpose, the authors developed a system consisting of bulbs attached to the ends of air condensers, with small tubes inserted into them to deliver oxygen to the sample. In 1942 Larsen et al. [50] developed a more advanced in-house method, which incorporated a thermostated oil bath for temperature regulation, a sintered glass thimble attached to a Pyrex tube for oxygen dispersion, a cold trap for preventing line blockages and water and carbon dioxide absorbers (see figure 2.6). This method was able to automatically register the oxygen consumed by a heated hydrocarbon.

Another system widely used in the bibliography is the Quartz Crystal Microbalance (QCM)/Parr bomb system coupled to a headspace oxygen sensor. The Parr bomb is a thermally regulated, 100 ml stainless steel vessel modified to vertically fit the QCM. This setup measures pressure, oxygen consumption, and deposit formation during the oxidation process [124–126]. Autoclave reactors have also been used for performing these tests. Alves-Fortunato et al. [58] reported the use of a 250 ml autoclave filled with 50 ml of sample. In these experiments, the autoclave is filled with argon, and heated to the goal temperature. Then argon

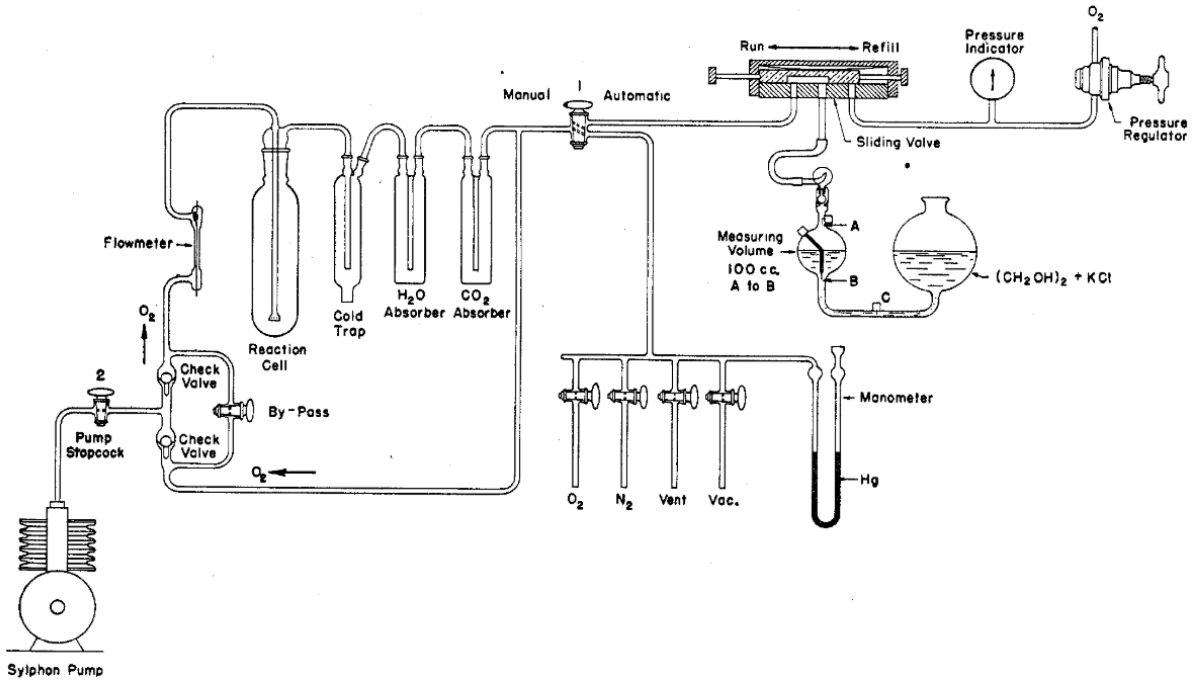


FIGURE 2.6 – Experimental setup used by Larsen et al. to perform accelerated oxidation experiments, reproduced from Larsen et al. [50].

is flushed and replaced with oxygen. This setup allows sampling of the reactor during analysis in order to follow the formation of oxidation products. More recently, Benrabah et al. [127] developed a microfluid reactor for the oxidation of liquid hydrocarbons. This system offers two key advantages: first, it simulates a perfectly stirred reactor, enabling the study of oxidation kinetics without interference from oxygen diffusion into the sample. Second, the transparent reactor allows real-time monitoring of the oxidation process using Raman spectroscopy. Despite advancements in accelerated oxidation tests, in-house developed methods often face reproducibility challenges. Additionally, variations in experimental setups make it difficult to compare results across studies. For the aforementioned reasons, the development of standardized methods is needed.

The Rancimat test is a standard method originally developed for the food industry, for testing edible oils and fats [128], which gained popularity due to its high reproducibility and ease of use [59]. Nevertheless, it has been extensively used in the energy industry for the analysis of lipids and Fatty Acid Methyl Esters (FAME) used in biodiesel production. The Rancimat test is described in detail in the EN 14112, in general, it works by detecting changes in conductivity; a sample is subjected to a constant flow of air (10 l/h) and heating at 120 °C, the air causes oxidizes the sample, causing the formation of polar products. The polar products are captured in a distilled water reservoir, where a conductimeter detects the maximum conductivity change (maximum of the second derivative). The time needed to reach this point is called Oil Stability Index (OSI) or Induction Period (IP) [17]. This method has been widely used to test biodiesel's stability, FAME, and oils [59, 128–132]. Some analyses have shown that

the Rancimat test results are correlated with other analytical parameters, such as the peroxide value, polymer content, acid value, kinetic viscosity, and ester content [128, 133].

The Jet Fuel Thermal Oxidation Stability Test (JFTOT) measures a fuel's tendency to form deposits when submitted to thermal stress. The method is described in the Standard Test Method for Thermal Oxidation Stability of Aviation Turbine Fuels, ASTM D3241. The apparatus utilized for this test consists of an electrically aluminum rod heated at 260 °C in contact with a constant flow of fuel for a duration of 2.5 h. The stability of the fuel is assessed according to two criteria; filter pressure drop less than 25 mmHg (or 3.3kPa), and visual examination of the deposits on the tube rated less than 3 [134]. The method has been widely used in the literature for the study of deposition mechanisms [48, 63], and the effect of copper and metal deactivators on deposit formation [135], while other studies focused on characterizing the deposits on the rod by metrological and chemical techniques, such as ellipsometry, laser scanning microscopy, Scanning Electron Microscopy (SEM), Fourier Transform Infrared Microscopy (FTIRM), Direct Analysis in Real Time Mass Spectrometry (DARTMS) and Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FTICRMS) [136]. However, JFTOT has been criticized for its lack of representation of real operating conditions which cause failures, such as servo valve sticking, heat exchanger degradation, and filter plugging [48]. Another of its main short-comings is the subjective evaluation of the deposit color, which is performed by visual comparison against a background of ASTM Standard Color Codes ranging from light tan to brown. When the deposit's color is different, such as white or peacock, it is considered as a test failure [134].

Another method is described in the Standard Test Method for Oxidation Stability of Spark Ignition Fuel–Rapid Small Scale Oxidation Test (RSSOT), ASTM D7525 [88], also known as the PetroOxy or RapidOxy test. This method tracks the oxygen consumption in a heated reactor during the oxidation of a sample. The Induction Period is defined as the time required to achieve a 10% pressure drop with respect to the maximum pressure [16]. This test has been used to determine the oxidation stability of diesel, biodiesel, and gasoline [58, 62], the effect of antioxidant addition [137], and the impact of branching, aromaticity and unsaturations on the stability of pure hydrocarbons [89, 91]. It has also been used in parallel with the Rancimat test [129, 138, 139] and the autoclave method [58]. In these studies, it was found that while there are some correlations between the results of these methods, the results are not comparable due to the different measurement principles, temperatures and oxygen partial pressures [129, 139, 140]. Some disadvantages of this method include uncertainties caused by changes in pressure due to reactions unrelated to primary oxidation. For instance, the bond cleavage may result in the formation of lower molecular mass compounds with low boiling point, resulting in an increment of pressure. On the other hand, the formation of polymers and deposits may result in a decrease of pressure [124, 125]. Furthermore, since the reactor cannot be sampled during the test, it is not possible to perform on-line analysis

during the measurement. Despite the aforementioned limitations, the RapidOxy method offers several advantages. Notably, it requires only a small sample volume, making it a cost-effective method. Additionally, the instrument enables measurements at higher temperatures compared to other devices, such as the Rancimat, resulting in shorter measurement times. Moreover, its extensive use in the literature facilitates the comparison of our findings with previously published results.

2.4 Materials and experimental method

2.4.1 Chemical reagents

For this work, we identified 95 high-purity and commercially-available hydrocarbons belonging to the hydrocarbon families and carbon number range of conventional and sustainable jet fuel [44, 141, 142]. The chemical products were acquired from Fischer Scientific® and Merck® catalogs, and were analyzed without further purification. A summary of the selected hydrocarbons is presented on table 2.1, while the full list of samples, along with their purity is provided in Appendix A. It is worth noting that various types of olefins were included in this study, despite their limited reporting in the literature, as they can significantly affect fuel stability even at low concentrations [89]. Olefin content is often not reported because they tend to coelute with naphthenes [44, 143]. However, estimates suggest that olefins make up approximately 1% of Jet-A1 and 6% of JP-5 [44].

Table 2.1 – Hydrocarbon families, carbon number ranges and numbers of samples in our database.

Empirical formula	Hydrocarbon family	Carbon number	Number of samples
n-C_nH_{2n+2}	*n*-paraffins	5-20	16
i-C_nH_{2n+2}	iso-paraffins	6, 8, 10, 12, 16	13
C_nH_{2n}	mono-naphthenes	5-10	11
C_nH_{2n}	mono-olefins	6, 8, 12, 16	10
C_nH_{2n-2}	di-naphthenes	10, 12	2
C_nH_{2n-2}	di-olefins	6-8	3
C_nH_{2n-2}	naphtheno-mono-olefins	5, 6, 9	3
C_nH_{2n-6}	mono-aromatics	6-15	26
C_nH_{2n-8}	naphtheno-mono-aromatics	9, 10, 12	4
C_nH_{2n-8}	mono-olefin-aromatics	9	1
C_nH_{2n-12}	di-aromatics	10-12	4
C_nH_{2n-14}	di-aromatics	12, 13	2
C_nH_{2n-14}	naphtheno-di-aromatics	12	1
C_nH_{2n-16}	naphtheno-di-aromatics	13	1

2.4.2 Instrumentation

The oxidation stability of the selected hydrocarbons was experimentally assessed using a RapidOxy 100 Fuel apparatus manufactured by Anton Paar® (see figure 2.7). The instrument consists of a pressure and temperature-regulated gold-plated chamber where the sample is placed. During the test, the chamber is filled with oxygen until a target pressure is reached, then it is heated to a set temperature [16, 97].

FIGURE 2.7 – RapidOxy 100 Fuel instrument and its gold-plated chamber. Reproduced from anton-paar.com/corp-en/products/details/rapidoxy-100.

At the start of the experiment, the system pressure increases due to the rising temperature, eventually reaching a maximum value, P_{max}. Under these conditions, the sample reacts with the oxygen present in the reactor, according to equation (2.3). Thus, the instrument measures the required time to register a 10% decrease of P_{max}, which is defined as the Induction Period (IP), as illustrated on figure 2.8. The Induction Period is considered to correspond to oxygen consumption caused by O_2 addition to the free radicals formed in the sample bulk [144].

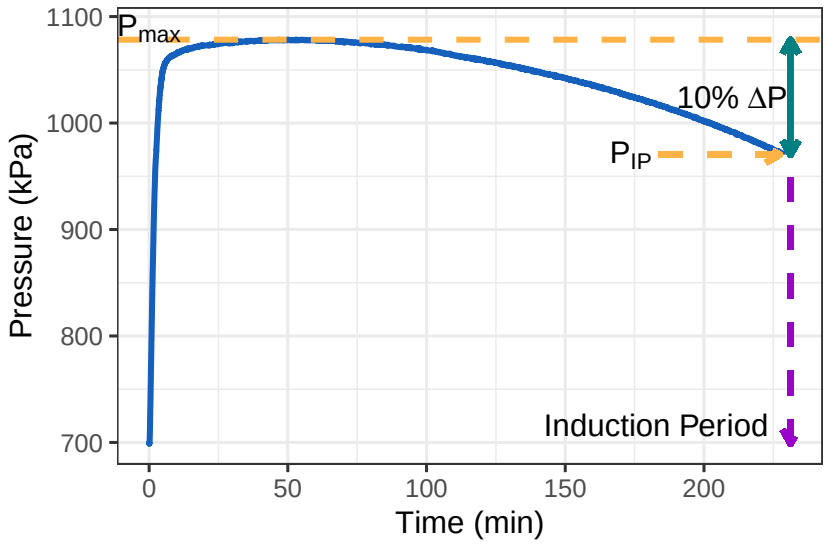


FIGURE 2.8 – Pressure vs. time curve obtained from the autoxidation of *n*-nonane at 140 °C, using a RapidOxy 100 fuel instrument with an initial oxygen pressure of 700 kPa.

The used experimental protocol was based on the ASTM D7545 [87]. First, we determined the appropriate sample volume. For liquid compounds, we adhered to ASTM D7545, which specifies a 5 ml sample. However, as the standard does not specify a quantity for solid samples, we employed an alternative strategy. First, we ensured the samples had melting points lower than the experiment temperature (140 °C). Then, we used the density of liquid-phase structure isomers for the calculation of the required mass for the analysis. The following formula was used:

$$m_{\text{solid}} = 5 \times \rho_{\text{liquid}} \tag{2.24}$$

where m_{solid} is the mass of the solid sample to be analyzed, 5 is the volume in ml used for the measurement of liquid samples, and ρ_{liquid}, the density of the isomer in g/ml. For example, 1,2,4,5-tetramethylbenzene ($C_{10}H_{14}$) is a solid at room temperature, while *n*-butylbenzene ($C_{10}H_{14}$) is a liquid. Given that the of *n*-butylbenzene is 0.86 g/ml, the required mass of 1,2,4,5-tetramethylbenzene is 4.3 g. This procedure ensures that the same amount of sample, in moles, is used for both compounds.

In the case isomers were not available, we used the density and molar mass of structurally similar compounds. In such instances, the required mass was calculated with the formula:

$$m_{\text{solid}} = 5 \times \rho_{\text{liquid}} \times \frac{MM_{\text{solid}}}{MM_{\text{liquid}}} \tag{2.25}$$

where m_{solid} is the mass of the solid sample to be analyzed, 5 is the volume used for measuring liquid samples, ρ_{liquid}, the density of the liquid, and MM_{solid} and, MM_{liquid} are

their respective molar masses. For instance, biphenyl ($C_{12}H_{10}$) is solid sample with molar mass(MM_{solid}) of 154.2 g/mol. On the other hand, the liquid compound with the most similar structure in our database is diphenylmethane ($C_{13}H_{12}$), which is a liquid with a density (ρ_{liquid}) of 1.01 g/ml and molar mass (MM_{liquid}) of 168.2 g/mol. Thus, the biphenyl mass required for the analysis is 4.63 g. Table 2.2 shows a list of the analyzed solid samples, the closest liquid-phase compound, and the weight used for analysis.

Table 2.2 – List of solid compounds in the database, detailing their chemical formulas, melting points, structurally similar isomers or compounds, and measured masses.

Solid compound	Formula	Melting point (°C)	Molar mass (g/mol)	Liquid compound	Formula	Density (g/ml)	Molar mass (g/mol)	Mass of solid sample (g)
1,2,4,5-tetramethylbenzene	$C_{10}H_{14}$	79.2	134.2	*n*-butylbenzene	$C_{10}H_{14}$	0.86	134,2	4.30
pentamethylbenzene	$C_{11}H_{16}$	54.4	148.2	*n*-pentylbenzene	$C_{11}H_{16}$	0.86	148.2	4.30
naphthalene	$C_{10}H_{8}$	80.3	128.2	tetralin	$C_{10}H_{12}$	0.97	132.2	4.70
2-methylnaphthalene	$C_{11}H_{10}$	35	142.2	1-methylnaphthalene	$C_{11}H_{10}$	1.00	142.2	5.00
biphenyl	$C_{12}H_{10}$	69.2	154.2	diphenylmethane	$C_{13}H_{12}$	1.01	168.2	4.63
fluorene	$C_{13}H_{10}$	116	166.2	diphenylmethane	$C_{13}H_{12}$	1.01	168.2	4.99
acenaphthene	$C_{12}H_{10}$	93.6	154.2	2-ethylnaphthalene	$C_{12}H_{12}$	0.99	156.2	4.89
1,4-di-*tert*-butylbenzene	$C_{14}H_{22}$	75-79	190.3	*n*-octylbenzene	$C_{14}H_{22}$	0.854	190.3	4.27

The rest of the experimental protocol is as follows. Before the beginning of the measurement, the instrument's chamber was purged by pressurizing to 700 kPa with oxygen and then, releasing the gas. After purging, the system was pressurized once again with oxygen to a final pressure of 700 kPa. In order to observe the effect of the temperature on the IP, experiments were carried out at in the range between 40 and 160 °C. After each measurement, the instrument's chamber and the injection nozzle were thoroughly cleaned with ethanol and dried with compressed air. All the measurements were performed by duplicate. To reduce measurement uncertainty, we considered only IP values greater than 20 minutes, as the instrument requires approximately four minutes to reach the target temperature.

2.5 Results and discussion

In this section, we discuss the measured Induction Periods for the analyzed compounds, based on their corresponding hydrocarbon families, i.e., paraffins, naphthenes, olefins, mono- and di-aromatics and naphtheno-aromatic hydrocarbons. The reported IP values were calculated as the mean value of replicates, and are reported at the standard temperature, 140 °C, unless otherwise stated. In certain cases, the measurement of the Induction Period was unattainable due to either the instrument's pressure limit or IP values being lower than the instrument's heating time. Given that the Induction Period follows the Arrhenius equation [62], we extrapolated the IP values determined at lower temperatures. These particular instances are denoted with an asterisk (*) hereafter. The list of IP for all the compounds analyzed in this study can be found in Appendix B.

2.5.1 Measurement precision

The first step of analysis result is the calculation of method precision. This stage is crucial for a fair comparison between sets of measurements, ensuring typical variability that can be expected during routine analysis instead of minimum variability. Precision is calculated from measurement repetitions performed under specified conditions. These specifications determine the levels of variability considered. Thus, there are three levels of precision [145, 146]:

- **Measurement repeatability:** Considers the lowest level of variability. It is obtained from measurements performed with the same measurement procedure, instrument system, analyst, and typically, the same day. Usually, it is the smallest possible variation, and is denoted as s_r.
- **Intermediate precision:** Also called within-lab reproducibility s_{RW}. It involves the variation typically found in a laboratory under longer periods of time, spanning several months. Some sources of variation includes different analysts, instrument systems, reagent batches. It is usually larger than the measurement repeatability.
- **Reproducibility:** Occasionally referred as between-lab reproducibility. It involves variations found in different laboratory settings.

In this work, we estimated the intermediate precision, s_{RW}, by considering measurements taken with two different RapidOxy 100 Fuel instruments. The acquisition of these measurements involved repeated analysis of the same sample over a time period ranging from weeks to months and, occasionally, the analysis of different sample batches. The calculation of s_{RW} was based on the pooled standard deviation, s_{pooled} [147, 148]. The pooled standard deviation can be understood as a weighted root mean square of the sample standard deviations. In the context of our work, a sample refers to each unique hydrocarbon-temperature combination. The formula for this statistic, as defined by Harris [147] is:

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2}{n_1 + n_2 + \ldots + n_k - k}}, \tag{2.26}$$

where:

- k is the number of samples,
- $s_1, s_2, \ldots, s_k$, the within-sample standard deviations, and
- $n_1, n_2, \ldots, n_k$ the number of parallel measurements for each individual sample.

In our work, the sum of n_i is 165, and $k = 81$. Another important factor to accurately determine the method precision is assessing whether the variance remains constant across the range of the measurand. For this purpose, we examined the measured Induction Period as a function of the mean Induction Period for all samples. As shown in figure 2.9**a**, the variance

increases with the magnitude of IP. This trend is further confirmed by plotting the absolute residuals against the mean IP, indicating heteroscedasticity due to the increasing variation across the IP range (see figure 2.9**b**). However, when plotting the relative residuals, as shown in figure 2.9**c**, the variance appears more consistent throughout the range. These results suggest that method precision should be reported as a relative value, by dividing s_{pooled} by the population mean. Given that the estimated s_{pooled} 0.75 h and the population mean is 6.42 h, the intermediate precision, expressed as the relative pooled standard deviation, is 11.7%. Our estimation of the method repeatability is congruent with the value of 10%, reported by Benrabah et al. [116].

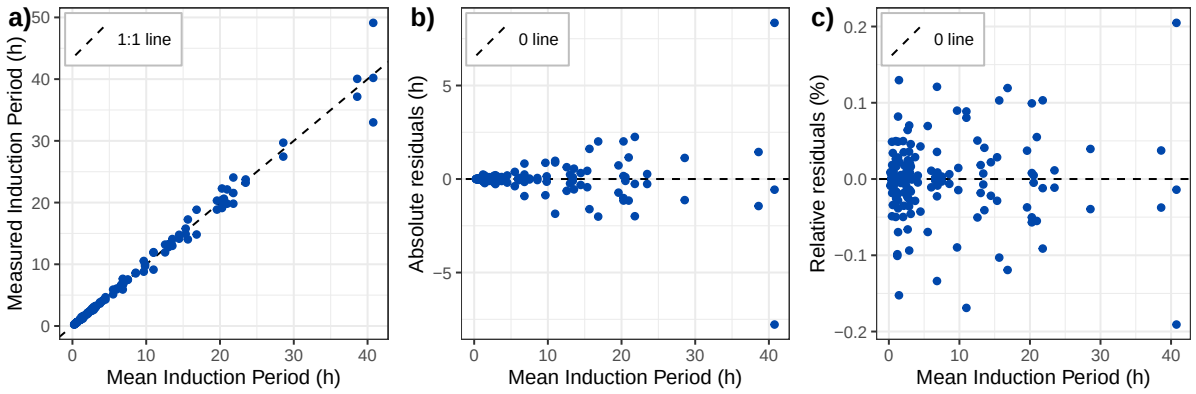


FIGURE 2.9 – Induction Period results for all the measured samples. **a)** Mean Induction Period vs. measured Induction Period. **b)** Mean Induction Period) vs. absolute residuals (mean Induction Period - measured Induction Period). **c)** Mean Induction Period vs. relative residuals.

2.5.2 Oxidation of paraffinic hydrocarbons

Parrafins or alkanes constitute the primary components of jet fuel by weight, having been extensively studied in the literature. Skolniak et al. [89] and Chatelain et al. [90] observed a non-linear dependence between the Induction Period and the chain length of linear alkanes. In this work, we expanded these results by measuring the Induction Period of all C_5-C_{20} linear alkanes. As shown in figure 2.10, the IP presents a sharp decrease when the carbon number increases from 5 to 10, reaching a plateau from C_{11} to C_{20}. *n*-tridecane presents a surprisingly high stability compared to *n*-dodecane and *n*-tetradecane, which could be caused by impurities present in the sample. Our results for C_{10}-C_{16} are congruent with the Induction Period values reported by Chatelain et al. [90]. However, the Induction Period of C_8 differs by 20%. Said difference may be explained by the presence of different chemical impurities in our reagents and different chemical providers.

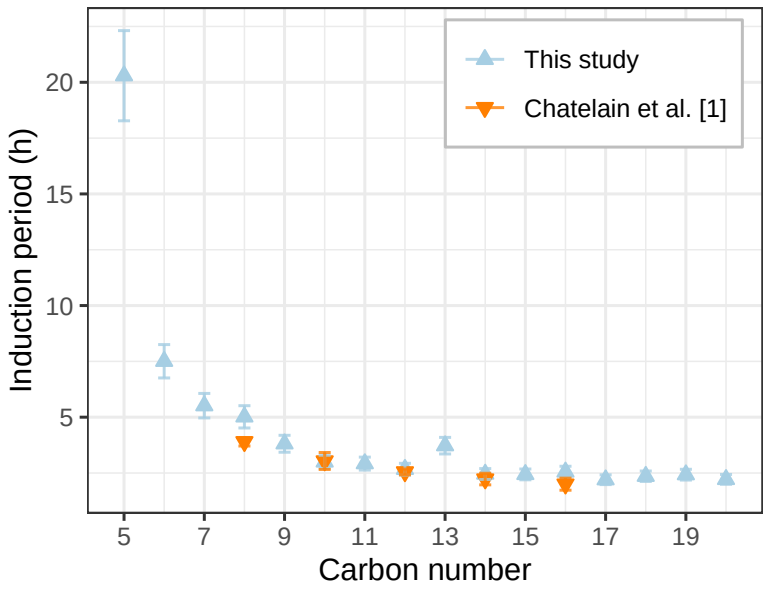


FIGURE 2.10 – Comparison of experimental Induction Periods for C_5-C_{20} *n*-alkanes at 140 °C between our study and Chatelain et al. [90] Uncertainties from Chatelain et al. [90] were reproduced from the original work accounting for reagent purity and instrument repeatability, expressed as $\Delta IP\ (h) = 0.015 IP_{measured}$.

The increase in reactivity observed for *n*-paraffins with increasing carbon number may be attributed to the occurrence of C-C bond cleavages in higher alkanes. Indeed, homolytic bond cleavage could lead to an increase in free radical concentration, subsequently elevating O_2 consumption. The results from Skolniak et al. [89] support this hypothesis. The authors reported that the main oxidation products of C_6-C_8 *n*-paraffins were ketones and alcohols with the same carbon number as the corresponding alkane, while the aging of *n*-hexadecane produced oxygenated compounds with lower carbon numbers. Some theoretical studies [149, 150] found that the C−C Bond-Dissociation Energy for the central C−C bond in linear C_6−C_{11} alkanes remains constant. However, Knyazev [151] reported that the per-bond rate constant values are positively correlated to paraffin chain length. The authors hypothesized that the increased C−C bond scission in linear alkanes with increasing chain length is due to torsional and bending motions that generate centrifugal forces on adjacent C−C bonds. This effect increments the pre-exponential factor in the Arrhenius equation, and consequently, the rate constant.

Branching effects play a significant role in the reactivity of iso-paraffins. For instance, Skolniak et al. [89] suggested that highly ramified paraffins were more stable than their linear counterparts due to steric hindering. However, Stark et al. [152] found that iso-paraffin reactivity was mainly related to carbon connectivity: tertiary hydrogen atoms being more reactive than secondary. Chatelain et al. [91] investigated the stability of various C_8 isomers. The authors expanded the findings of Stark et al. [152], demonstrating that compounds containing

quaternary carbons tend to exhibit greater stability compared to their linear and less ramified counterparts.

In the present study, we found that both the carbon connectivity and the quantity of substituted C centers impact the oxidation stability. Induction periods of various C_6 isomers are given in figure 2.11. For instance, 2,3-dimethylbutane exhibits an Induction Period approximately 90% lower than *n*-hexane's, and about 80% lower than the two mono-branched isomers. Thus, the reactivity of a hydrocarbon is related to the number of tertiary C atoms. Additionally, our findings suggest that branching position also influences oxidation stability, as evidenced by 2-methylpentane presenting an Induction Period approximately 28% lower than that of 3-methylpentane. Our results align with the findings of Hudzik et al. [153], who reported that the Bond-Dissociation Energy of tertiary carbon atoms depended on their position in the chain. For instance, for 2,3-dimethylpentane and 2,2,3,4-tetramethylpentane, the authors calculated that the tertiary carbon's C-H BDE is up to 1 kcal/mol higher in the 3- position with respect to the 2-position. However, the analysis of similar isomer pairs may be needed to corroborate this reactivity trend.

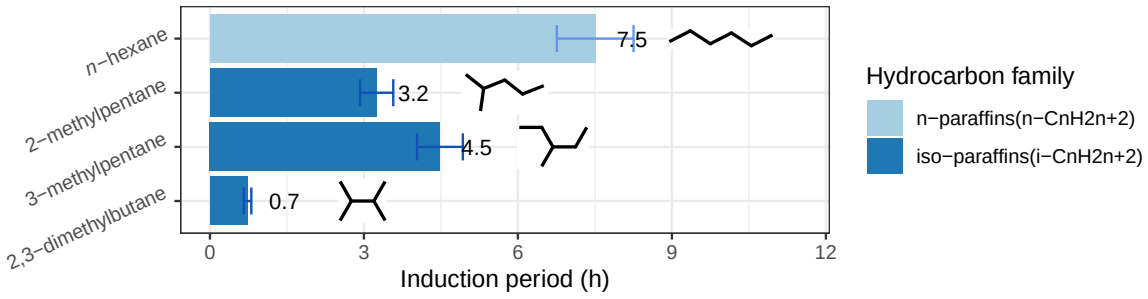


FIGURE 2.11 – Experimental Induction Period for C_6 alkane isomers at 140 °C.

A previous study performed by Chatelain et al. [91] reported that iso-paraffin reactivity is correlated to carbon type available in the molecule; quaternary C containing molecules being the most stable. However, our results show that quaternary C may negatively affect hydrocarbon's oxidation stability. As illustrated in figure 2.12, the ramified C_8 isomer is 5.4 times more stable than *n*-octane; however, the branched C_{12} is only 50% more stable than *n*-dodecane, while 2,2,4,4,6,8,8-heptamethylnonane is twice more reactive than *n*-hexadecane. One potential explanation is that at low number of quaternary carbons, the lack of H atoms causes and stabilizing effect on the molecule. However as the molecule size increases, C-C bonds with at least one quaternary carbon atom may be more likely to break and form free radicals compared to C-C bonds between secondary carbon atoms in linear alkanes. As reported by Zhu et al. [154], alkane C-C Bond-Dissociation Energy lowers with the number of ramifications. For instance, the C-C BDE for secondary carbons in C_8 isomers is approximately 87.2 kcal/mol, while the C-C BDE for two contiguous quaternary carbons in 2,2,3,3-tetramethylbutane and 2,2,3,3,4,4-hexamethylhexane is 81.0 kcal/mol and 52.8 kcal/mol, respectively.

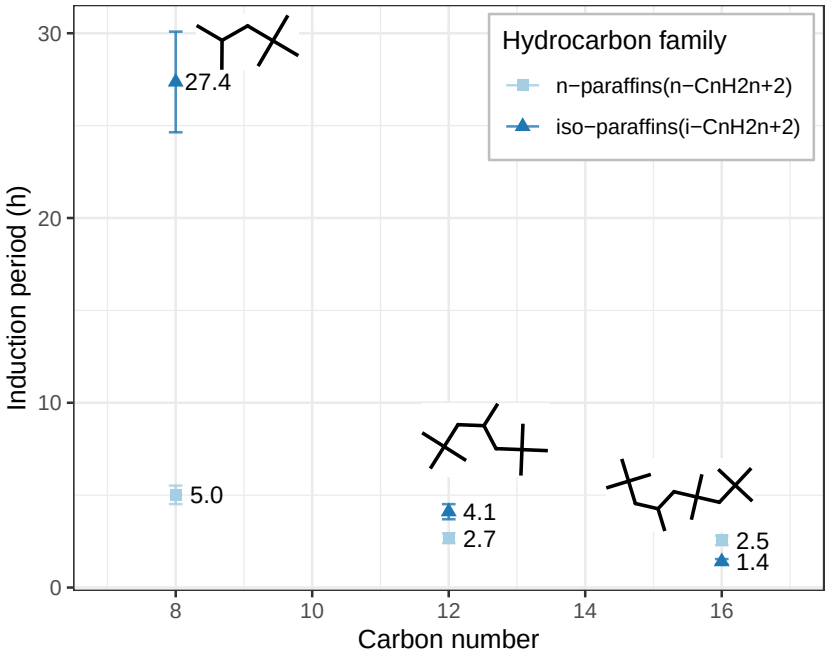


FIGURE 2.12 – Experimental Induction Period for C_8-C_{16} linear and branched paraffins at 140 °C.

2.5.3 Oxidation of naphthenes, alkylnaphthenes, and di-naphthenes

While naphthenes or cycloalkanes present a similar structure to alkanes, their reactivity differs. For instance, figure 2.13**a** shows that unsubstituted C_5-C_8 naphthenes are more reactive than their linear counterparts, except for cyclohexane. Notably, cyclohexane presents the highest stability, with an Induction Period comparable to that of *n*-hexane. In contrast, other naphthenes become progressively more reactive as their carbon number deviates from six and the ring strain increases. This is better visualized when observing the relationship between ring strain and the Induction Period. Figure 2.13**b** shows a negative correlation between the Induction Period of unsubstituted naphthenes and ring strain. Our results agree with the findings of Agapito et al. [155]. In their work, the authors found that ring strain destabilization lowers the C-H Bond-Dissociation Energy of cyclopentane, with respect to cyclohexane. Some studies identified and quantified the oxidation products of C_5-C_8 naphthenes [89, 117]. In broad terms, hydroperoxide content increases with temperature until it plateaus and subsequently declines. At this point, the concentration of secondary oxidation products, such as ketones, alcohols, and dicarboxylic acids, increases.

In the case of substituted cycloalkanes, H-abstraction can occur at the ring's secondary and tertiary positions, or at the chain's primary, secondary, or tertiary positions. Figure 2.14 shows that the number of substituents in the ring has a negative effect on the stability. For instance, compared to the IP of cyclohexane (6.8 h), methylcyclohexane presents an IP decrease of ≈ 70% (2.1 h), 1,3-dimethylcyclohexane ≈ 75% (1.8 h) and 1,2,4-trimethylcyclohexane ≈ 90% (0.6 h). However, the chain length does not seem to have a relevant effect on the stability

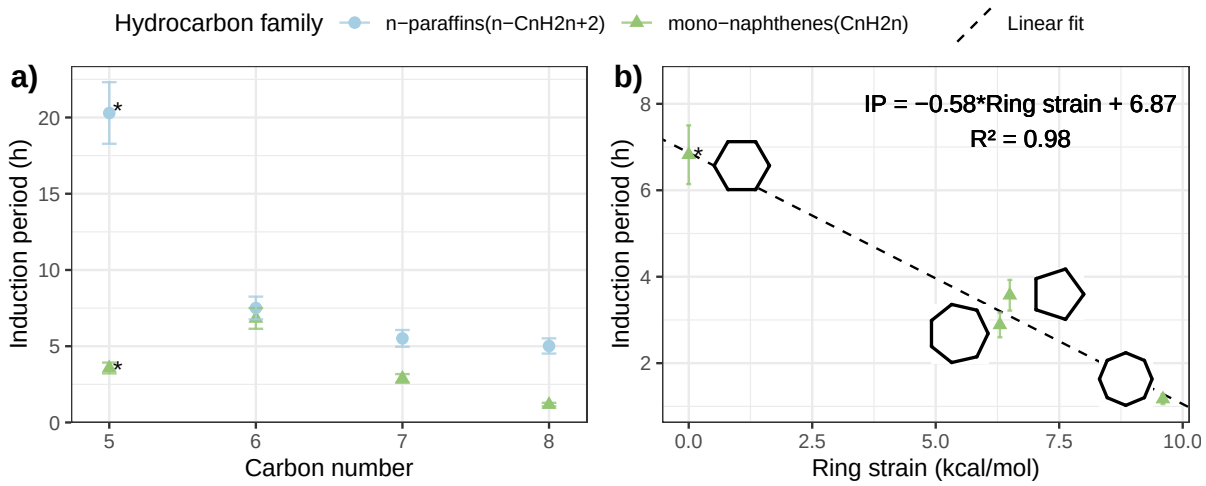


FIGURE 2.13 – **a)** Experimental Induction Period for unsubstituted C_5-C_8 naphthenes and their corresponding *n*-parrafins at 140 °C. **b)** Induction period as a function of ring strain for C_5-C_8 unsubstituted naphthenes. *Cyclopentane's IP was extrapolated from measurements performed at 110, 120 and 130 °C by plotting log(IP) vs. 1/T (K).

since methyl- to *n*-butylcyclohexane have comparable IP values. These results suggest that reactivity of alkyl derivatives of cyclohexane, is mainly related to H-abstraction at the ring's tertiary and secondary sites.

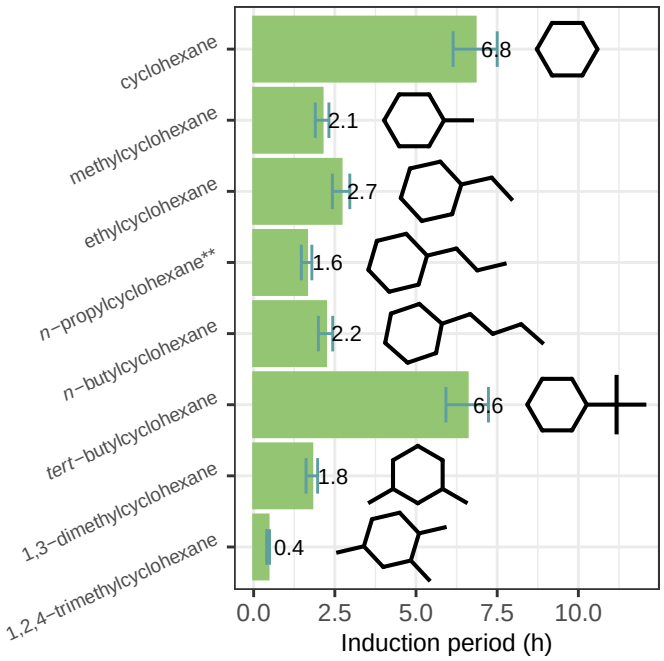


FIGURE 2.14 – Experimental Induction Period for alkyl derivatives of cyclohexane with different alkyl chain length and bonding patterns at 140 °C. **Obtained from Ben Amara et al. [16].

Our findings are consistent with those of Lian et al. [156], who showed that, for *n*-propyl and *n*-butylcyclohexane, the hydrogen abstraction rate constants for secondary carbons in the naphthene ring are higher than those for secondary carbons in the substituting chain.

Furthermore, the authors showed that the tertiary C in the ring has a similar kinetic constant to that of the secondary C in the ring. However the branching ratio suggested that H-abstraction mostly occurs at the tertiary site. Additional evidence for this hypothesis is the high stability of *tert*-butylcyclohexane, which suggests that the tertiary ring site cannot steadily undergo H-abstraction due to steric hindering caused by the *tert*-butyl group, thus H-abstraction occurs at the secondary positions in the ring, providing a similar stability to that of cyclohexane.

Di-naphthenes exhibit significantly higher reactivity compared to mono-naphthenes. As shown in figure 2.15, decalin has an Induction Period nearly seven times shorter than that of cyclohexane, while the Induction Period of bicyclohexyl is approximately half as long. This increased reactivity is primarily due to the presence of tertiary carbon atoms in these compounds. Jaffe et al. [157] found that, although H-abstraction can occur at all sites in decalin, the primary oxidation products predominantly result from hydroperoxide formation at tertiary sites. In contrast, data on the autoxidation products of bicyclohexyl are scarce. However, its comparatively higher stability may be linked to the formation of stable compounds such as cyclohexane and cyclohexene following bond cleavage at tertiary sites [158].

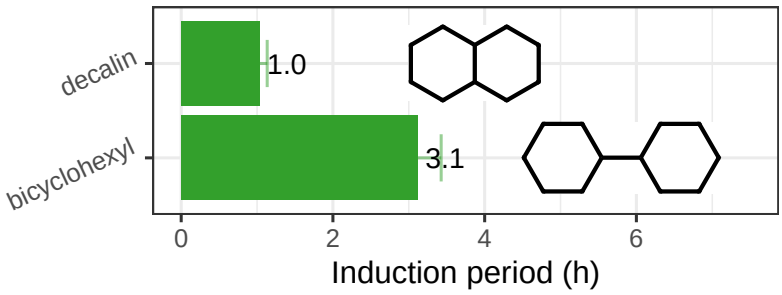


Figure 2.15 – Experimental Induction Period for di-naphthenes at 140 °C.

2.5.4 Oxidation of olefinic hydrocarbons

Olefins or alkenes are generally more reactive than their saturated counterparts. Such difference in reactivity could be caused by the low Bond-Dissociation Energy of allylic H atoms and the formation of resonance-stabilized allylic radicals [159].

In figure 2.16, the Induction Periods of several types of olefins are presented. The results support that the stability of olefins is related to the olefin substitution pattern: monosubstituted (1-hexene) > disubstituted cis (cyclohexene) > disubstituted trans (*trans*-2-hexene), > disubstituted geminal (2,3-dimethyl-1-butene) > tetrasubstituted (2,3-dimethyl-2-butene), an opposite reactivity trend to the well-known enthalpy of hydrogenation in alkenes. This trend could be explained by the fact that highly substituted alkenes have a higher number of allylic H atoms; for instance, 1-hexene has 2, while 2,3-dimethyl-2-butene has 12. While technically, 2,4,4-trimethyl-2-pentene is a trisubstituted olefin, its IP is comparable to that of a disubstituted

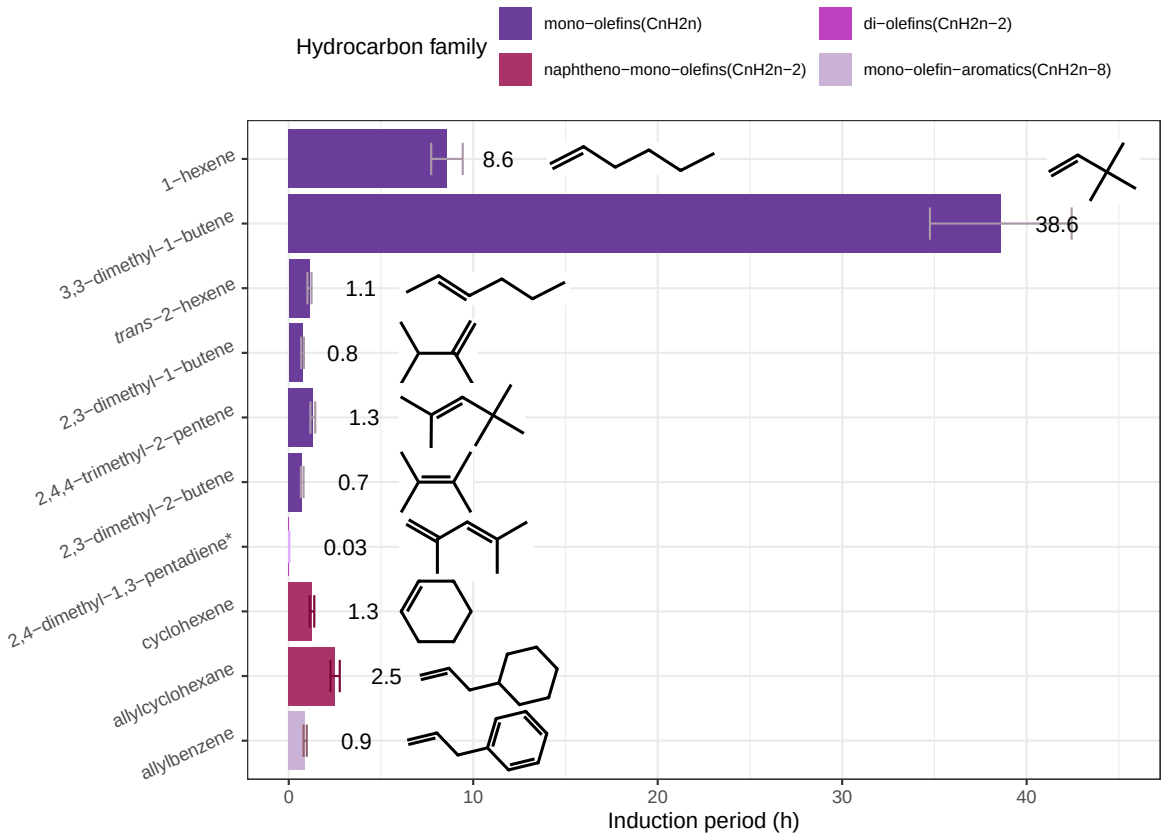


FIGURE 2.16 – Experimental Induction Period for alkenes at 100 °C. *2,4-dimethyl-1,3-pentadiene's IP was extrapolated from measurements performed at 40 and 60 °C by plotting log(IP) vs. 1/T (K).

olefin. This may result from the presence of a tertiary group at the allylic position, which influences reactivity in two ways: the absence of an H atom and the hindrance of H-abstraction at the geminal allylic position. Additionally, 3,3-dimethyl-1-butene exhibits greater stability than 1-hexene despite both being monosubstituted alkenes. This behavior may be attributed to the absence of allylic H atoms in 3,3-dimethyl-1-butene, further supporting the hypothesis that H-abstraction occurs at the allylic sites. However, its lower stability compared to its saturated counterpart, 2,2-dimethylbutane (IP = 171.2 h at 100 °C), suggests that oxidation proceeds through at least one additional mechanism, such as O_2 addition at the double bond.

The low stability of allylbenzene and 2,4-dimethyl-1,3-pentadiene supports the hypothesis that olefins' reactivity is related to the formation of delocalized radicals. In the case of allylbenzene, its double bond is not conjugated with the aromatic ring; however, the radical formed after H-abstraction at the allylic site becomes delocalized throughout the entire structure. Further evidence supporting this theory is the higher stability of allylcyclohexane compared to allylbenzene's, whose structure does not present delocalization after H-abstraction. On the other hand, the high reactivity of 2,4-dimethyl-1,3-pentadiene could be explained by the presence of 9 allylic H atoms, which produce delocalized radicals in the two conjugated double bonds.

Non-conjugated di-olefins are less reactive than conjugated dienes, since the radicals formed after H-abstraction at the allylic sites are not stabilized by the conjugated system. Intuitively, an increase in the number of double bonds would be associated with higher reactivity. For instance, as shown in figure 2.17, 1,7-octadiene reacts 5 times faster than 1-octene. However, 1,5-hexadiene is 1.6 times more stable than 1-hexene. Although detailed studies on the primary oxidation of 1,5-hexadiene are lacking in the literature, some proposed mechanisms for its thermal degradation indicate that, at low temperatures, 1,5-hexadiene undergoes internal rearrangements and cyclization, forming stable molecules like benzene. These findings may help explain the high stability of 1,5-hexadiene [160–162].

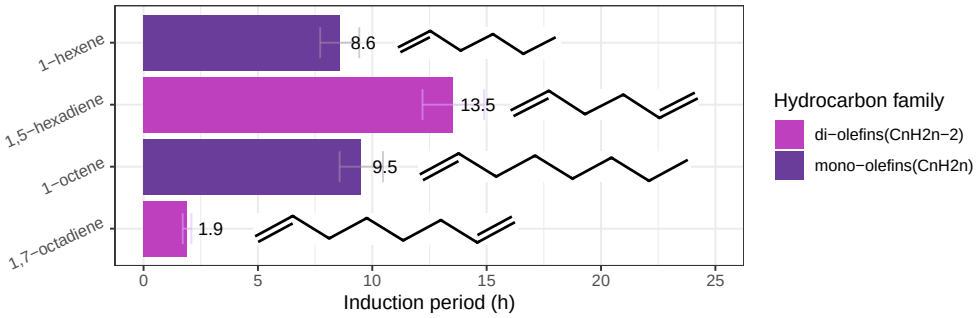


FIGURE 2.17 – Experimental Induction Period for C_6 and C_8 mono- and di-olefins at 100 °C.

Contrary to linear alkanes, the chain length does not have an impact on olefin reactivity. As shown in figure 2.18, the Induction Period of 1-alkenes from C_6 to C_{16} remains constant, despite the great difference in allylic to secondary H atom ratio. This finding suggests that the allylic sites outcompete other sites during H-abstraction.

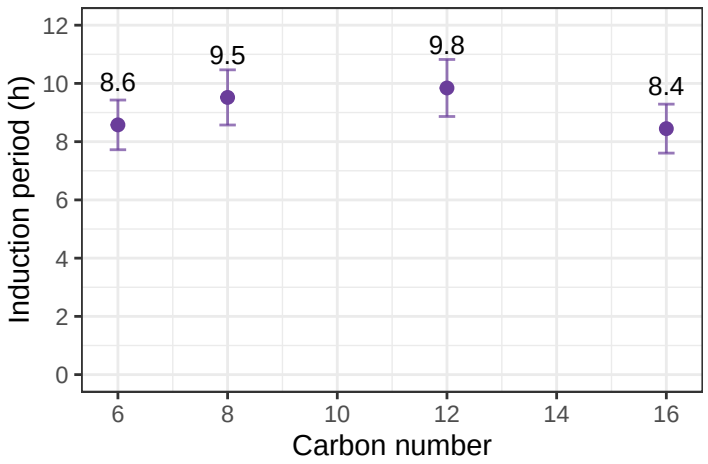


FIGURE 2.18 – Experimental Induction Period for C_6-C_{16} 1-alkenes at 100 °C.

In the case of cyclo-alkenes, ring strain seems to have a significant role on the oxidation stability. As shown in figure 2.19, cyclopentene, a molecule with a strained ring, is 5 times more reactive than cyclohexene.

A previous study proposed that olefins mainly react through two mechanisms, abstraction at the allylic site, and addition to the vinylic positions. The first mechanism seems to be dominant reaction pathway, since mono-peroxides were found at higher amounts than polyperoxides [163]. However, the addition mechanism, and further decomposition of the intermediate products explain the presence of *tert*-butanol in the oxidation of 3,3-dimethyl-1-butene. [164, 165] Thus, the main products of olefin oxidation are hydroperoxides, epoxides, carbonyl compounds resulting from the cleavage of C=C bonds, and polyperoxides.

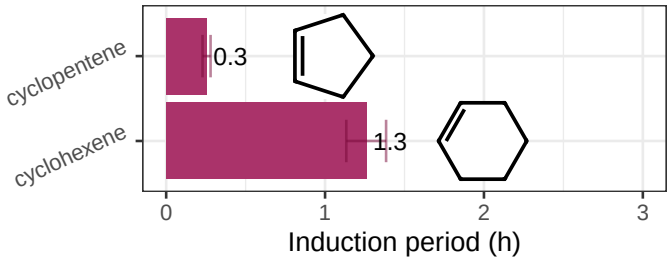


FIGURE 2.19 – Experimental Induction Period for cyclopentene and cyclohexene at 100 °C.

2.5.5 Oxidation of aromatic and alkylaromatic compounds

The effect of the length of the aliphatic chains attached to aromatic rings has been previously studied by Larsen et al. [50], who studied the oxidation of *n*-pentyl, *n*-hexadecyl, and *n*-octadecylbenzene, and observed that the chain length had a negative effect on the compounds' stability. A similar trend was highlighted by Ben Amara et al. [16], who found that the Induction Period of *n*-propylbenzene was 50% lower than that of toluene.

In our study, we observed trends consistent with those previously reported in the literature. Figure 2.20 portrays the relationship between the Induction Period and the length of the paraffinic chain attached to a phenyl group. It is worth noting that, to the best of our knowledge, we report for the first time the IP of benzene (20.3 h) at standard conditions (temperature = 140 °C, initial pressure = 700 kPa, sample volume = 5 ml). In comparison, the IP value reported by Skolniak et al. [89] is 15.7 h. This value was obtained at 140 °C, with an initial pressure of 500 kPa and a sample volume of 10 ml [89]. From our results, two different behaviors can be observed; a slow and steady decline in stability for aromatics with a chain length of less than 3, and a sharp decline followed by a plateau for aromatics with a chain length greater than 3. In the first region, benzene and toluene present nearly identical IP values. This finding seems to contradict chemical intuition, since benzene is expected to react slower than toluene, due to the high C-H BDE of phenylic sites (472 kJ/mol) compared to toluene's benzylic site (375 kJ/mol). Additionally, the H-abstraction in benzene results in the formation of an unstable phenyl radical, in contrast with the stable benzyl radical formed during the oxidation of toluene [166]. In the case of ethylbenzene, it presents an IP approximately 20% lower than toluene's, which

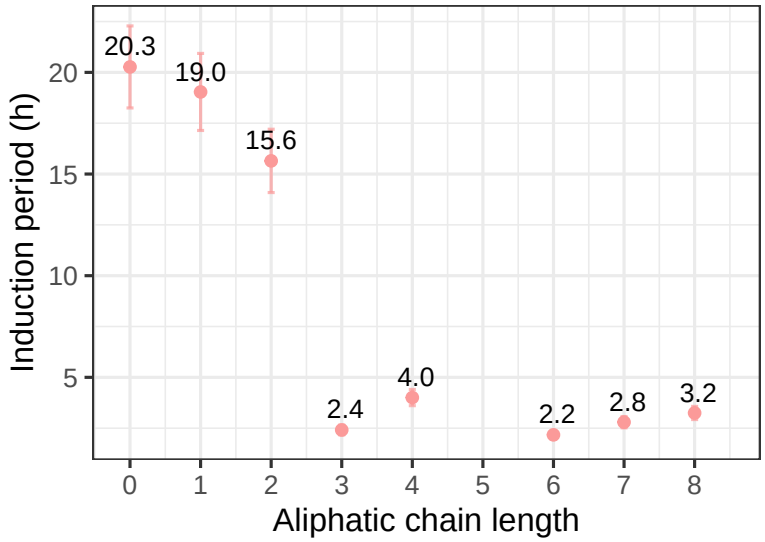


FIGURE 2.20 – Experimental Induction Period for benzene and *n*-alkylaromatic compounds at 140 °C.

could be explained by the presence of secondary benzylic carbons which are more reactive than primary carbons. In the second region, *n*-propylbenzene and the rest of aromatic compounds present an increased reactivity compared to ethylbenzene, toluene and benzene, which could be explained by two factors. The first is the presence of reactive secondary benzylic carbons. And the second, high amount of secondary carbons, which easily undergo H-abstraction.

Although the length of the substituting chain significantly influences the stability of alkylaromatics, it appears that the bonding pattern of the benzylic C atom plays a more substantial role [50, 121]. As shown in figure 2.21, the reactivity of alkylaromatics may be mainly related to the degree of substitution at the benzylic site. In the case of butylbenzene isomers, *sec*-butylbenzene, possesses a tertiary benzylic carbon and exhibits greater reactivity compared to *n*-butylbenzene, while *tert*-butylbenzene presenting a quaternary benzylic carbon and no further H to abstract, presents a much higher stability compared to the other isomers. The same trend is found in propylbenzene isomers: cumene is more reactive than *n*-propylbenzene. However, in the case of isobutylbenzene, which features a tertiary carbon not positioned in the benzylic site, it presents an Induction Period comparable to the one of *n*-butylbenzene, providing additional evidence of the reactivity of benzylic sites. Thus, the stability of alkylaromatics with respect to the benzylic carbon bonding pattern is as follows: quaternary > secondary > tertiary.

Ben Amara et al. [16] investigated the effect of the number of substitutions on the oxidation stability of aromatic hydrocarbons by comparing the induction period of benzene, toluene, *m*-xylene, and 1,2,4-trimethylbenzene. The authors concluded that there wasn't a significant correlation between the two variables. However, as shown in figure 2.22, the number of substituents may have an important effect on the reactivity depending on their relative position. The Induction Periods of benzene and toluene are similar to those of *m*-xylene, *p*-xylene and

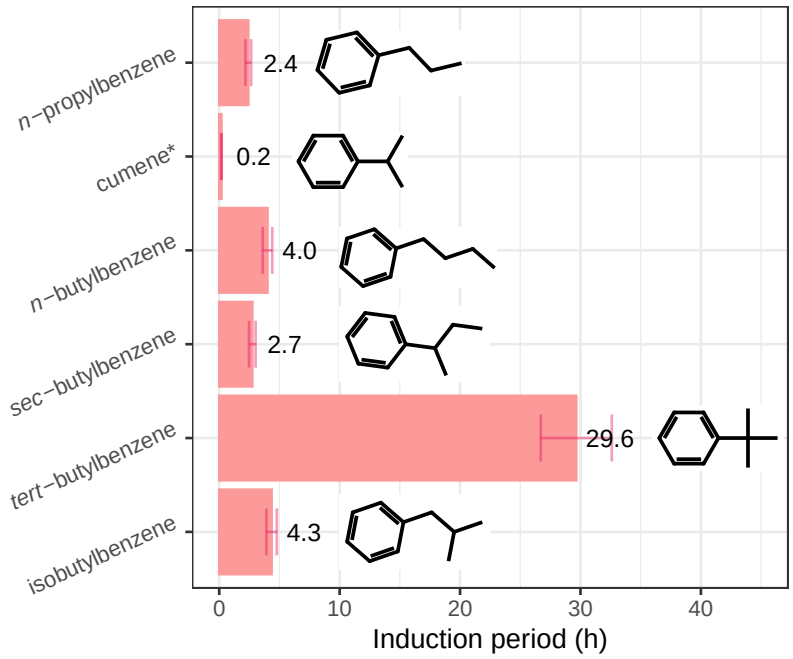


FIGURE 2.21 – Experimental Induction Period for different alkylaromatics with different connectivity patterns at 140 °C. *Cumene's IP was extrapolated from measurements performed at 100 and 120 °C, by plotting log(IP) vs. 1/T (K).

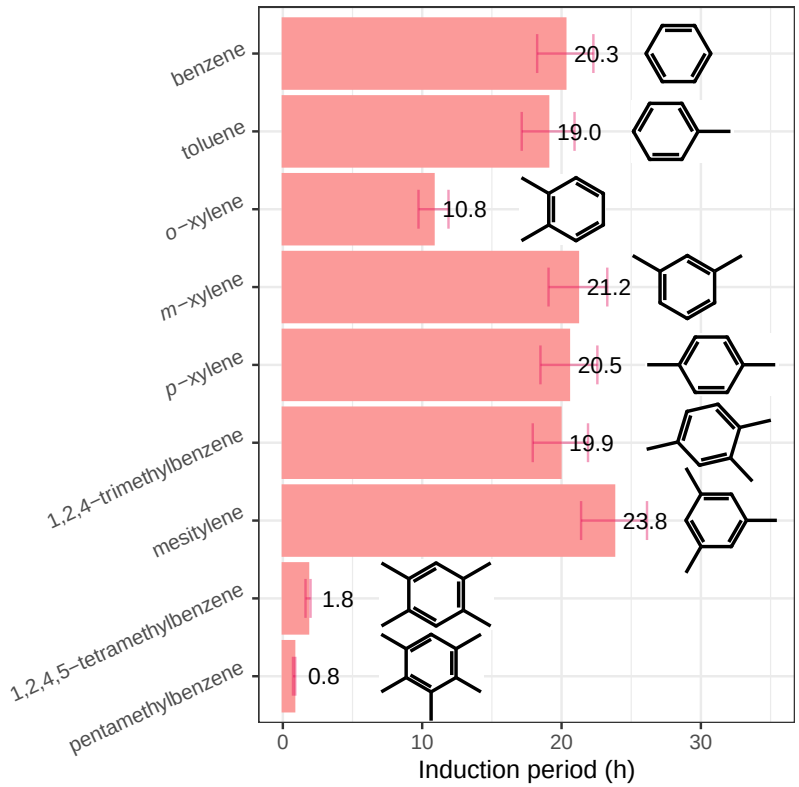


FIGURE 2.22 – Experimental Induction Period for benzene and aromatics with 1 to 5 methyl groups at 140 °C.

trimethylbenzene isomers; however, *o*-xylene presents an IP 50% lower than its positional isomers. This reactivity trend was observed by Meziane et al. [167] in their study using a jet-stirred reactor. This behavior could be explained by the mechanism developed by Kukkadapu et al. [168], which shows that *o*-xylene forms phthalan after undergoing H-abstraction, O_2 addition, and a final internal isomerization step that produces a hydroxyl radical and facilitates chain propagation reactions. This reaction mechanism might explain the slight reactivity difference between 1,2,4-trimethylbenzene and mesitylene. However, the stability contrast between 1,2,4-trimethylbenzene and *o*-xylene could be attributed to H-abstraction competing at both the ortho- and para- positions in 1,2,4-trimethylbenzene. Finally, 1,2,4,5-tetramethylbenzene (durene) and pentamethylbenzene present a reduction in their Induction Period of 90% and 95%, respectively, compared to toluene. This behavior could be explained by two factors: the abundance of methyl groups, which facilitate H-abstraction, and their positioning at ortho-sites.

The number of substitutions on the aromatic ring also affects reactivity when the benzylic carbons are secondary or tertiary. As shown in figure 2.23, the stability of ethylbenzene decreases by approximately 80%, while the stability of cumene decreases by 28%. This could be explained by the increase of the number of reactive sites in the molecules. However, 1,3,5-triisopropylbenzene is more stable than 1,4-diisopropylbenzene, seemingly contradicting this trend. We hypothesize that this may be caused by impurities present in the sample, given the lower purity of the reagent (95%). For compounds with quaternary carbons, 1.4-di-*tert*-butylbenzene is more stable than *tert*-butylbenzene. This increased stability may be due to the non-reactive nature of quaternary benzylic carbons, along with the steric hindrance from bulky *tert*-butyl groups that impedes reactions at the phenylic sites.

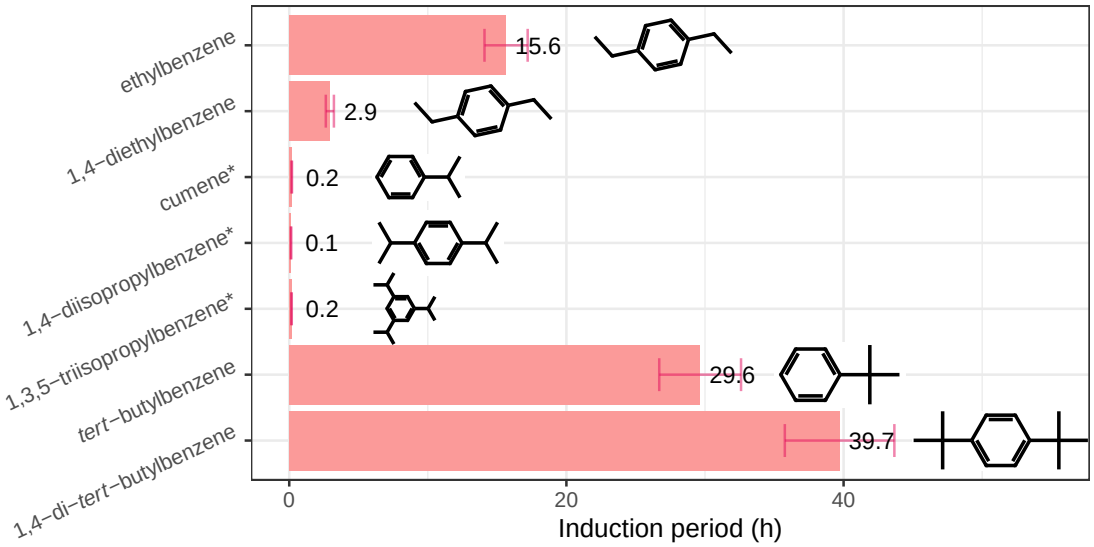


FIGURE 2.23 – Relationships between the number of substitutions and the Induction Period for alkylaromatics with secondary, tertiary and quaternary carbon atoms at the benzylic sites. Measurements were conducted at 140 °C. *The IP values of cumene, 1,4-diisopropylbenzene and 1,3,5-triisopropylbenzene were extrapolated from measurements conducted at lower temperatures: 100 and 120 °C for cumene, and 80, 100 and 120 °C for 1,4-diisopropylbenzene, and 80 and 100 °C for 1,3,5-triisopropylbenzene. This was achieved by plotting log(IP) against 1/T (K).

2.5.6 Oxidation of di-aromatic and naphtheno-aromatic hydrocarbons

As previously described in section 2.5.5, the stability of alkylaromatics is closely related to the number of substituents, their connectivity, and their relative position. Similarly, the stability of di-aromatics is mainly determined by these factors. As depicted in figure 2.24, biphenyl exhibits the highest stability. Due to its absence of substituents, H-abstraction must occur at the phenylic sites, a process hampered by its high Bond-Dissociation Energy and the formation of the phenyl radical, which, unlike the benzyl radical, is not stabilized by resonance. On the other hand, diphenylmethane has a benzylic site activated by two aromatic rings, allowing for resonance stabilization in both rings. Fluorene has a structure similar to diphenylmethane; however, the extra ring results in a nearly planar structure that increases its aromaticity, which could account for its higher reactivity compared to diphenylmethane.

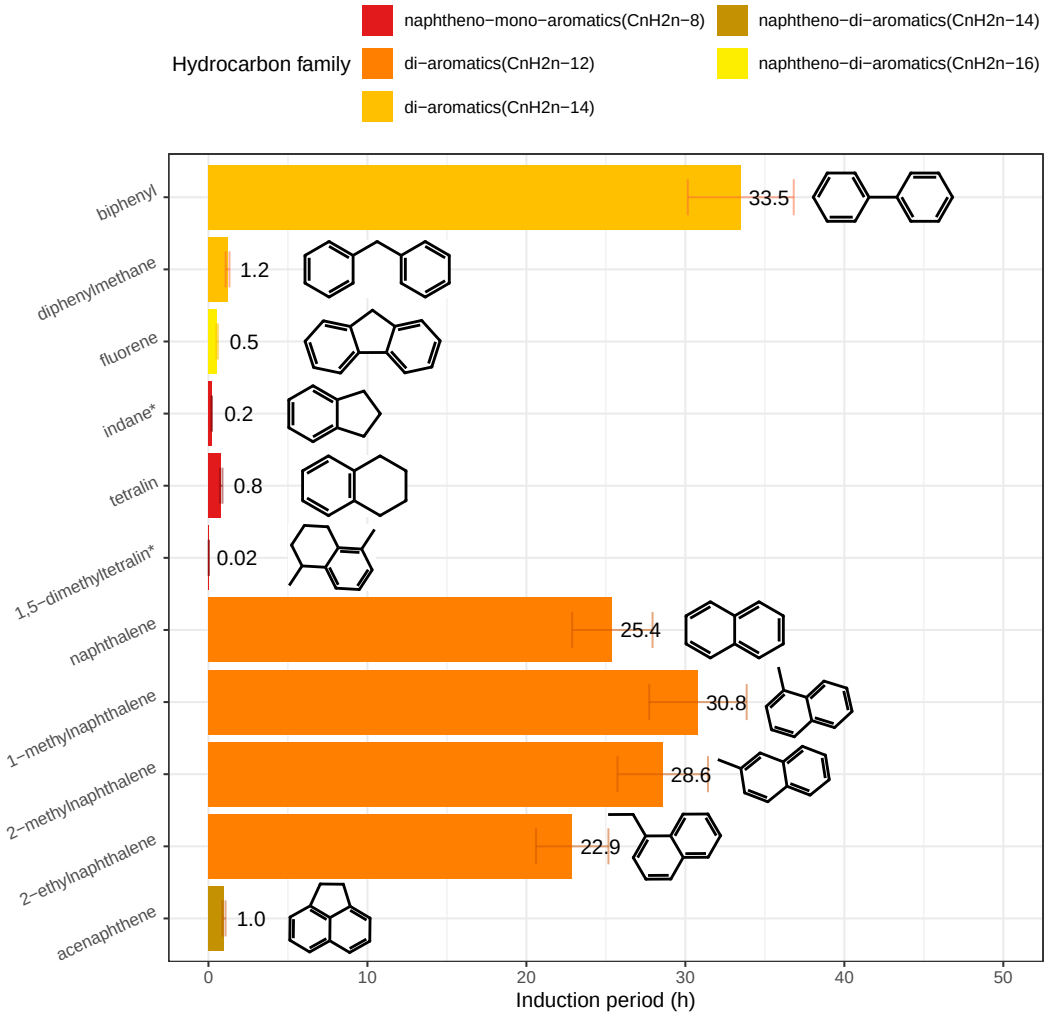


FIGURE 2.24 – Experimental Induction Period for naphtheno-aromatic and di-aromatic hydrocarbons at 140 °C. *The IP values of indane and 1,5-dimethyltetralin were extrapolated from measurements conducted at lower temperatures: 100 and 120 °C for indane, and 100 and 80 °C for 1,5-dimethyltetralin. This was achieved by plotting log(IP) against 1/T (K).

Naphtheno-aromatics, or partially hydrogenated di-aromatics like indane and tetralin, exhibit shorter Induction Periods than naphthalene. The reactivity of these two compounds may be attributed to the four benzylic hydrogens activated by the aromatic rings. However, indane's higher reactivity compared to tetralin could be due to ring strain destabilization, as discussed in section 2.5.3. Like alkylaromatics, the reactivity of naphtheno-aromatics is influenced by the bonding at the benzylic site; for example, 1,5-dimethyltetralin is 40 times more reactive than tetralin due to the presence of a tertiary carbon.

In the case of fused di-aromatics, we found that naphthalene presents a lower IP compared to that of 1- and 2-methylnaphthalene. This contradicts the reactivity trend of alkylaromatic compounds, where the presence of benzylic carbons primarily determines their reactivity. Nevertheless, our results are inconclusive due to the purity difference between naphthalene (99%) and the methylnaphthalene isomers (96%), since previous studies have shown that the

IP is affected by the reagent's purity [90]. To potentially resolve this ambiguity, product purifications could be carried out in the future, but this is beyond the scope of this thesis.

Another reactivity trend discussed in the bibliography, is the effect of substituent position on the oxidation stability of naphthalene derivatives. Larsen et al. [50] and Shaddix et al. [169] observed that the 1-position is more reactive than the 2-position in methylnaphthalene isomers. Similarly to alkylaromatics, the substituent's chain length negatively impacts the stability of naphthalenes, as 2-ethylnaphthalene has a lower induction period than both methylnaphthalene isomers. Like indane and tetralin, acenaphthene exhibits enhanced reactivity compared to naphthalene, 1-methylnaphthalene, and 2-methylnaphthalene, attributable to the presence of secondary benzylic sites within its strained ring structure. However, acenaphthene is twice as stable as fluorene, despite having four benzylic H compared to the two of fluorene. We hypothesize that the observed reactivity trend may be attributed to the dual activation of fluorene's benzylic carbon, in contrast to acenaphthene, where the reactive sites are activated by only one aromatic ring.

2.5.7 Temperature effect on the Induction Period

The Induction Period presents an exponential decay with respect to the experiment temperature. For instance, Bacha et al. [62] observed that a temperature reduction of 10 K resulted in roughly doubling the IP. This exponential relationship can be mathematically represented by the equation:

$$\mathrm{IP}\,(T) = \alpha \exp(\beta T)\,, \tag{2.27}$$

where T is the temperature, α is a pre-exponential factor, which can be interpreted as the IP at a reference temperature or $T = 0$, and β is the rate of change with temperature. We can define a "decay factor", γ, which is the value that scales the IP when the temperature is increased by 10 °C or 10 K. For example, for the results of Bacha et al. [62], $\gamma = 2$. The generalized form of this equation is given by:

$$\mathrm{IP}\,(T + 10) = \frac{1}{\gamma}\mathrm{IP}\,(T)\,. \tag{2.28}$$

After substituting the general form of the exponential decay equation (2.27), in equation (2.28), we obtain:

$$\alpha \exp\left(\beta(T + 10)\right) = \frac{1}{\gamma}\alpha \exp(\beta T). \tag{2.29}$$

Therefore, the expression for γ can be obtained by dividing equation (2.29) by $\alpha \exp(\beta T)$ and rearranging the resulting expression:

$$\gamma = \exp(-10\beta). \tag{2.30}$$

Then, β can be estimated by performing a linear fit between the $(\log(\mathrm{IP}_i), T_i)$ pairs for a sample. This is easily demonstrated by applying the logarithm to both sides of equation equation (2.27), resulting in:

$$\log(\mathrm{IP}) = \log(\alpha) + \beta T, \tag{2.31}$$

where the constant term represents the logarithm of the pre-exponential factor α, while the slope corresponds to β.

In this work, we performed a systematic analysis of the decay factor, γ for the analyzed compounds, shown in figure 2.25. It is worth noting that some compounds, such the decay factor of some compounds, such as 1,2,4,5-tetramethylbenzene and fluorene are not presented, since the measurement of solid samples could cause damage to RapidOxy instrument.

In figure 2.25 we present the decay factor γ for the measured compounds. The mean decay factor $\bar{\gamma}$ for all the analyzed hydrocarbon families is approximately two, which is consistent with the results of Bacha et al. [62]. However some deviations from the mean value can be observed. The decay factor of linear paraffins, shown in figure 2.25**a**, is almost constant. The most notable exception is *n*-octane, whose γ value is approximately 2.75. The reactivity of this family is mainly related to the reactive methylene groups that increase with the carbon number, however the odd behavior of *n*-octane may be explained by the presence of impurities in the sample. On the other hand, branched paraffins, in figure 2.25**b**, present a lower decay factor compared to their linear counterparts. Nevertheless, γ seems to be positively correlated with the molecule carbon number, increasing from 1.6 for C_6, to 2.1 for C_{16}. This suggests that iso-paraffins are more susceptible to the effect of temperature as they increase in size, possible due to the increase of methylene and methine groups.

Naphthenes (figure 2.25**c**) present a mean decay factor 2.2, similar to that of linear paraffins. Some notable deviations from the mean value are cyclopentane, cyclooctane, 1,2,4-trimethylcyclohexane and bicyclohexyl. The higher temperature susceptibility of cyclopentane and cyclooctane may stem from the tendency of strained cycloalkanes to undergo ring-opening reactions at elevated temperatures [117]. In contrast, 1,2,4-trimethylcyclohexane might be less temperature-sensitive, as its reactivity is primarily linked to the abundance of tertiary carbon atoms rather than ring-opening reactions. For bicyclohexyl, increased reactivity with temperature may be attributed to potential cleavage of its two-ring structure. Olefins (figure 2.25**d**) have a mean γ value of 2.1, with 2,3-dimethyl-2-butene and allylbenzene being notable outliers,

exhibiting γ values of approximately 1.2 and 1.4, respectively. This behavior may be attributed to the ability of both compounds to form free radicals that are stabilized by two or more double bonds. On the other hand, the average γ of aromatics is 1.7, the lowest among the hydrocarbon families. As shown in (figure 2.25**e**), two sets of compounds can be observed. The first are compounds with high γ values, which contain long aliphatic chains, or tertiary benzylic carbon atoms. The latter are compounds with low γ values, whose reactivity is linked to phenylic sites since they don't contain benzylic hydrogen atoms, such as benzene and *tert*-butylbenzene, as well as molecules that contain methyl groups attached to aromatic rings, such as toluene, *m*-xylene and mesitylene. Interestingly, compounds with methyl groups in ortho- positions, such as *o*-xylene and 1,2,4-trimethylbenzene are more susceptible to temperature effects, which could be related to their specific reactivity discussed in section 2.5.5. Finally, diaromatics (figure 2.25**f**) present a $\bar{\gamma}$ of 1.8, with acenaphthene having a notably higher decay factor.

The analysis of the decay factors provides another perspective for the study of oxidation stability. Not only it is important to focus on high Induction Period values, but also, on the sensibility of a sample to temperature. For instance, a compound A may be more stable than a compound B at a given temperature, however, this trend could be reverse at a different temperature. Thus, the formulation of fuels should consider both the Induction Period and the decay factor, γ.

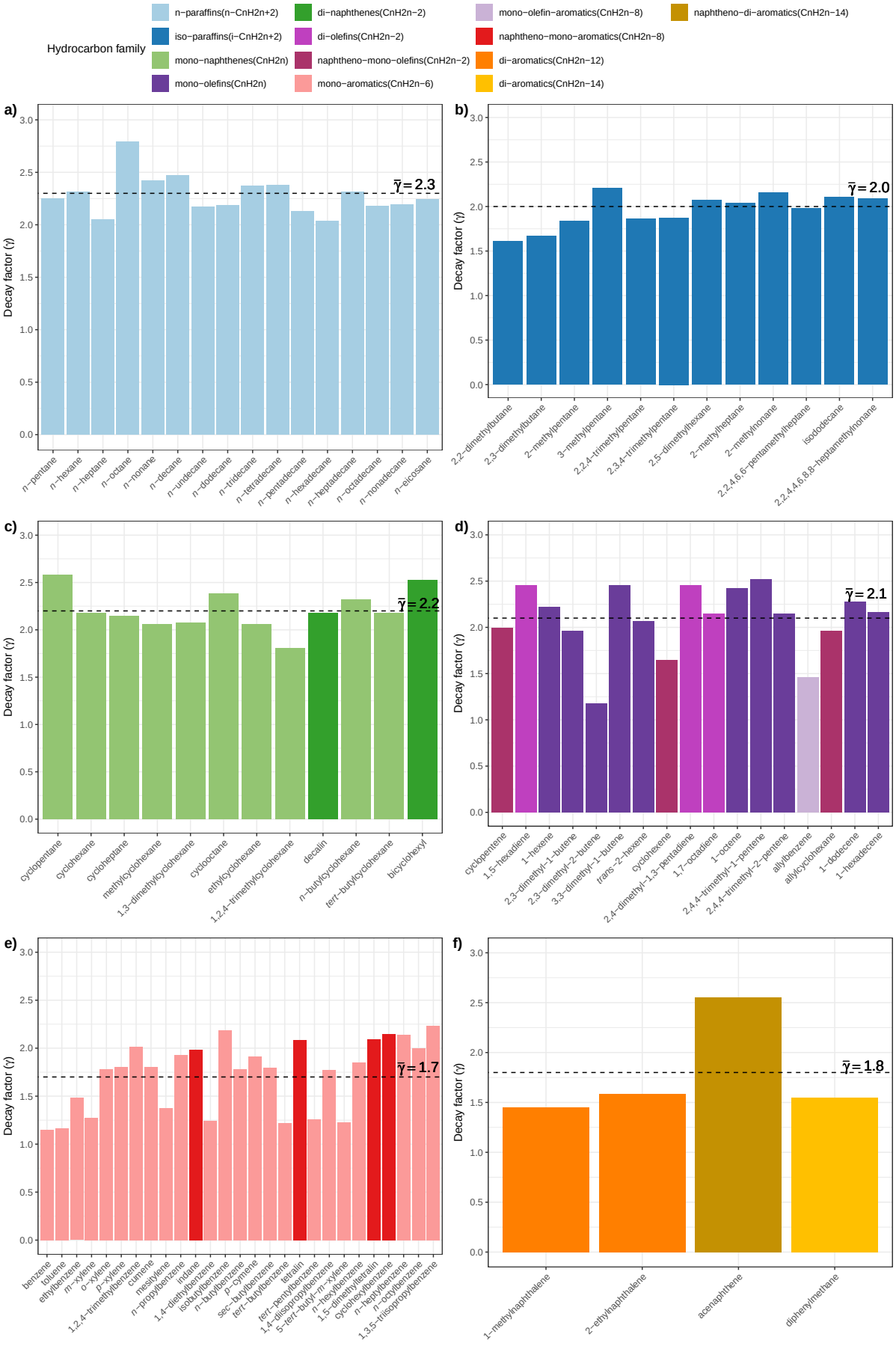


FIGURE 2.25 – Decay factor γ for the analyzed compounds.

2.6 Conclusions

In this study we assessed the oxidation stability of hydrocarbons belonging to different chemical families, such as linear and branched paraffins, naphthenes, olefins, aromatics, di-aromatics and partially hydrogenated di-aromatics by using the RapidOxy instrument to measure their Induction Period. A systematic analysis of the results allowed us to expand previous studies and to identify new structure-property relationships.

For *n*-alkanes, we complemented bibliography data [90] by reporting the Induction Period at 140 °C of C_6-C_{20} compounds. In the case of branched paraffins, we observed two new trends: firstly, the branch position in methylpentane isomers impacts the stability; and secondly, the presence of quaternary carbons can negatively impact the stability of molecules, especially at higher carbon numbers. This may be due to the reduction in the C-C Bond-Dissociation Energy between two contiguous ramified carbon atoms.

On the other hand, we observed that the stability of naphthenes presents a non-linear trend with respect to the ring size, which could be caused by strain effects. Thus, cyclohexane has the highest stability, while cyclopentane, cycloheptane and cyclooctane are more reactive. Alkylnaphthenes present lower stability than naphthenes, however their reactivity seems to be related to the tertiary sites in the substituted rings and not to the chain-length. In the case of olefins, we observed that their high reactivity is related to the formation of resonance-stabilized allyl radicals, thus a reaction trend opposite to the hydrogen enthalpy was observed. Consequently, we found that olefins lacking allylic hydrogen atoms are 1-2 orders of magnitude more stable than other isomers of equal carbon number. Furthermore, we noted that delocalization caused by conjugated double bonds and aromatic rings favors H-abstraction, while the chain length has a negligible impact on the stability. Non-conjugated olefins do not present a clear trend in reactivity; while they present more reactive allylic sites, these compounds may undergo isomerization and cyclization reactions, forming stable compounds such as benzene. Similarly to naphthenes, the stability of cyclo-olefins is also influenced by ring destabilization effects.

In the case of aromatic compounds, contrary to the reactivity trend reported by Ben Amara et al. [16], we found that the number of substituents does negatively impact the stability and furthermore, methyl groups in ortho- position are more reactive than when in meta- and para-positions. Additionally, we noted that chain length is negatively correlated with the IP, and that C connectivity at the benzylic site has a great influence in the reactivity, tertiary C-containing molecules being more reactive than alkylaromatics with linear chains, and quaternary C-containing molecules being 7.5 times more stable than the linear counterparts. The stability of unfused di-aromatics is greatly affected by the presence of doubly activated benzylic sites and ring strain. In the case of naphthalenes, substituent position didn't seem have a significant effect contrary to previously reported data [50]. Nevertheless, we observed that the chain

length of the substituent has a small negative effect on the compound's stability, while ring strain greatly lowers the stability of fused aromatics.

We calculated the decay factor for the compounds analyzed, finding an average decay factor of approximately 2. This indicates that a 10 °C increase in the accelerated oxidation experiment results in a halving of the Induction Period. However, certain compounds, such as benzene, toluene, and *m*-xylene, showed significant deviations from this average value. Analyzing decay factors provides additional insight into oxidation stability, highlighting that hydrocarbon selection for fuel formulation should consider both oxidation stability and temperature sensitivity, as indicated by the decay factor.

Finally, this work represents a first experimental step in the development of predictive approaches. We have identified several molecular features that impact the stability of compounds when exposed to an oxidizing environment, and highlighted various interesting trends as a function of the molecular structure. All this information now needs to be supplemented and streamlined. For instance, by performing new IP measurements and using machine learning algorithms to derive models based on quantitative structure-property relationships that could accurately predict the stability of any hydrocarbon and, at a later stage, formulate hydrocarbon mixtures with properties close to real fuel.

Chapter 3

QSPR-based modeling of the oxidation stability of hydrocarbons

3.1 Introduction

In the previous chapter we discussed the fundamentals of the oxidation phenomenon; the reaction mechanism, molecular features related to the oxidation stability, and the experimental methods used to characterize this property. Furthermore, we presented the results from our study, where we oxidized a wide range of hydrocarbons relevant to jet fuel, using the PetroOxy/RapidOxy accelerated oxidation test. Now, we will discuss modeling as an alternative approach for the study of autoxidation.

As discussed in section 1.4, the oxidation phenomenon is typically studied using kinetic, or "white-box" modeling. However, this approach relies on mechanisms developed for the gas phase, which must be corrected to account for solvent effects in the liquid phase. These corrections present a significant challenge, especially given the numerous parameters that need to be determined for complex matrices like fuels.

Conversely, data-driven modeling offers an alternative to first-principles approaches. It relies on data collection to establish relationships between input and output variables, without requiring prior mechanistic knowledge. Machine-learning algorithms can further assist in uncovering these relationships, making data-driven modeling a powerful tool for the study of complex systems.

In this chapter, we present a data-driven modeling study based on the experimental results discussed in the previous chapter. To provide context, we briefly examine kinetic modeling, highlighting its limitations and explaining our preference for black-box modeling approaches. Then, we introduce key concepts related to data-driven modeling, machine-learning, model validation, and Quantitative Structure-Property Relationship (QSPR). Next, we present our

modeling results, compare the performance of different algorithms, and conclude by interpreting the best-performing model, identifying the most relevant features for the prediction of the Induction Period.

3.2 Data-driven modeling: Cheminformatics

In contrast with analytical modeling, data-driven modeling can establish relationships between input and output variables without requiring prior knowledge of a system's internal workings (see figure 3.1). In chemistry, data-driven approaches have proven valuable for analyzing large datasets generated by analytical techniques such as spectroscopy, chromatography, and kinetics [170]. Black-box modeling employs multivariate techniques to analyze spectra and chromatograms, which involve hundreds to thousands of variables due to the discretization of instrument signals. For kinetic modeling, data-driven approaches permit the study of complex phenomena, such as autoxidation, without the need to explicitly account for the hundreds to thousands of chemical species and reactions typically involved in the process.

FIGURE 3.1 – Black-box modeling: Inputs are converted into outputs through a multivariate function, without knowledge of the internal workings.

Data-driven modeling for chemistry applications is broadly applied in two scientific fields: cheminformatics and chemometrics. This chapter focuses on the former, while the latter will be discussed in the next chapter. One of the earliest definitions of cheminformatics was proposed by Brown [171]. Paraphrasing his definition, "cheminformatics is the mixing of information technology and management to transform data into information, and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization". However, modern definitions do not consider that cheminformatics is solely concerned about drug discovery [170]. For instance, Varnek and Baskin [172] define cheminformatics as a theoretical chemistry discipline complementary to quantum chemistry and force-field molecular modeling, which is based on the representation of molecules as objects (graphs or vectors) in a chemical space.

Cheminformatics incorporates several fields, including the development of chemical data bases, exploration of chemical space, and Quantitative Structure-Activity Relationship (QSAR)/ Quantitative Structure-Property Relationship (QSPR) models. In this work, we discuss the latter. Such modeling develops a mathematical relationship between a chemical response and quantitative chemical attributes that encode molecular structural features (see figure 3.2). This

modeling approach receives its name from the studied chemical response, for instance, biological activity (QSAR), a physicochemical property (QSPR), toxicity (QSTR), biodegradability (QSBR), among others [173, 174]. In this work, we will refer to this family of methods as QSPR.

FIGURE 3.2 – QSPR modeling expressed as a black-box approach.

3.2.1 Quantitative Structure-Property Relationships (QSPR)

Quantitative Structure-Property Relationship (QSPR) is based on the similarity principle, which states that compounds with similar molecular structure exhibit similar physicochemical properties. In general, it consists in identifying relations between properties of interest and molecular structures, which chemical descriptors can describe [175, 176]. These relations have the general form:

$$\text{property} = f(\text{descriptors}). \tag{3.1}$$

Once a correlation between a structure and a property has been found, it can be used to screen any number of chemical compounds, including those not yet synthesized to identify compounds with the desired properties [177].

An important limitation of QSPR modeling is the violation of the similarity principle, often referred to as "activity cliffs". The concept of activity cliffs is closely related to the "activity landscape", which represents an N-dimensional space comprising the molecular representation of a group of compounds and their associated target property. According to the similarity principle, the activity landscape is envisioned as a meadow-like terrain, where changes in the target property occur smoothly over small distances. However, activity cliffs, as the term suggests, represent abrupt changes in the target property within a short distance. These cliffs can lead to significant mispredictions in regions where the overall model predictivity is otherwise high [178].

Another limitation of QSPR is the accurate modeling of trace properties. While bulk properties are primarily determined by the structure of the chemical compounds present in large quantities, trace properties are influenced by compounds at very low concentrations [179]. This poses a challenge for modeling oxidation stability, as it can be affected by the presence of dissolved metallic ions or antioxidant species [17].

3.2.1.1 Molecular representation

Molecular representation is the process of converting molecular information into an object through a well-specified algorithm [180, 181]. In this work, we discuss three types of molecular representation: linear notations, graphs and descriptor vectors. Linear notations use alphanumeric characters to encode a molecule's structure, allowing it to be easily communicated to a computer. One of the most widely used notation systems is Simplified Molecular Input Entry System (SMILES). SMILES notation uses ASCII characters to easily encode a molecular structure, as shown in table 3.1.

Table 3.1 – SMILES notation for different C_6 hydrocarbons.

Molecule	**SMILES**
n-hexane	CCCCCC
2-methylpentane	C(C)CCCC
1-hexene	C=CCCCC
cyclohexane	C1CCCCC1
benzene	c1ccccc1

Besides SMILES, other notation systems exist. For instance, in 2006, the International Union of Pure and Applied Chemistry (IUPAC) implemented the International Chemical Identifier (InChI), a system where each molecule has an unique 27-character identifier [182]. Linear notations are valuable because computational software can convert these strings into a 2D molecular graph or a 3D molecular structure [182]. For instance, SMILES Arbitrary Target Specification (SMARTS) is an extension of SMILES that serves as a language for specifying substructures using a defined set of rules. Thus, a particular molecular substructure expressed as a SMARTS pattern can be found in a molecular graph [183].

A graph is an abstract structure consisting of nodes that are connected by edges [182]. In the case of molecular graphs, atoms are represented as nodes, while bonds are represented as edges (see figure 3.3. In some instances, a node may represent a substructure instead of an atom [172]. A property can be calculated from a molecular graph or an adjacency matrix associated to it, this is also referred as a graph-level task [184].

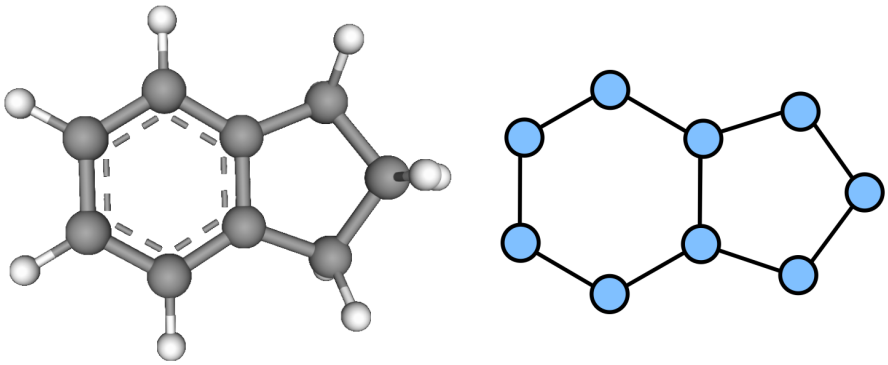


FIGURE 3.3 – 3D molecular structure and molecular graph of indane.

Descriptor vectors are mathematical entities that encode a molecular structure (see figure 3.4). These molecular representations are also called “fingerprints”. There are different types of fingerprints, including dictionary-based and circular fingerprints [185]. Dictionary-based fingerprints represent functional groups by counting the occurrences of predefined substructures in a molecule [180]. Examples of fingerprints in this category include PubChem fingerprints, Molecular ACCess System (MACCS), and Barnard Chemistry Information (BCI) fingerprints, among others [185]. In contrast, circular fingerprints are not based on predefined fragments and do not have fixed lengths. They are generated by centering on a non-hydrogen atom and iteratively extending to its neighbors until all fragments of the molecule are listed or a custom limit is reached. Notable examples include Extended Connectivity FingerPrints (ECFPs) and Functional-Class FingerPrints (FCFPs), among others [185]. No single fingerprint is optimal for all applications. Combining multiple fingerprint types can improve the description of a molecule [180]. For instance, Sandfort et al. [186] combined 24 fingerprints into a 71 375-dimensional vector.

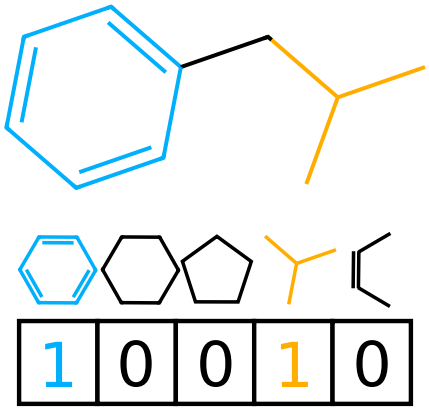


FIGURE 3.4 – Vector representation of isobutylbenzene, encoding the presence or absence of a molecular fragment.

In general, the space spawned by descriptor vectors is easier to navigate than the topology space formed by molecular graphs. Thus, the study of vector-space can be performed with

widely used multivariate statistic techniques, such as Principal Component Analysis (PCA) and hierarchical cluster analysis [172].

Molecular representations can encode different levels of information according to the structural features they consider. The descriptors calculated from such representations can be classified as follows [177, 187–191]:

- **Experimental descriptors**: Physicochemical properties that are obtained experimentally, such as molar refractivity, enthalpy of formation (ΔH_f), partition coefficient (P), among others [187, 192].
- **Theoretical descriptors**: Derived from the representation of molecules, as structural or empirical formulas. They can be classified in 5 types according to their information content and ease of calculation:
  - **0D descriptors:** Based on chemical formulas, the simplest representation of molecules. These descriptors account for the occurrence of elements in a molecules, they are easy to calculate, but provide little information and have high degeneracy (i.e. a descriptor has the same value for different molecules). Some examples include atom counts, molecular weight and properties than can be calculated as sum or average of atomic contributions, such as atomic van der Waals volumes [188, 192]
  - **1D descriptors:** Based on structural formulas, these descriptors increase the provided information by accounting for the presence of substructures, like functional groups and atom-centered fragments, e.g., primary carbons, secondary carbons, etc [188, 192].
  - **2D descriptors:** Also known as topological indices, these descriptors account for the connectivity and adjacency of atoms. 2D descriptors are calculated from molecular graphs, where atoms and bonds are represented as vertices and edges, respectively. They include information about atom connectivity, and are sensible to molecule size, shape, symmetry, branching and cyclicity. Topological indices can be furthered classified into (1) topostructural indices, which only consider information from atom adjacency and through-bond distances, and (2) topochemical indices, which also encode chemical properties, such as hybridization state. Some examples include the Zagrep index, the Kappa shape index, Wiener Balaban J index, among others. [187–189, 192].
  - **3D descriptors:** Based on the 3D structure of molecules and the spatial coordinates of atoms. This class includes geometrical descriptors, such as interatomic distances, dihedral angles, valence angles, molecular volume, 3D-Molecular Representation of Structures based on Electron diffraction (MoRSE), GEometry, Topology, and Atom-Weights AssemblY (GETAWAY) and surface area, and quantum descriptors, such as atomic charges, electronic features, vibrational frequency levels, reactivity indices, HOMO and LUMO energies, among others. Their main disadvantages are their

high computational cost, dependence on the geometry optimization method, and their inability to account for different conformations in highly flexible molecules [177, 187, 189, 192].

– **4D descriptors:** These descriptors account for the interactions between molecules and active sites of biological receptors. The 4th dimension can be added to a regular 3D structure by including different conformations, alignments, orientations and protonation states. Other approaches can be obtained by using a probe to map the surface of a molecule, located on an equally spaced, 3D grid [190–192].

Ideally, molecular descriptors should offer a clear structural interpretation, effectively distinguish between isomers, demonstrate good predictive performance for the target property, and have a balance between complexity and molecular resolution [188].

In the last decade, QSPR has been used to predict several properties of hydrocarbons typically found on jet fuel, diesel, biodiesel and gasoline [175, 193]. For example, Saldana et al. [193] used a type of vector representations, known as Functional Group Count Descriptors (FGCD) to encode the structural information of hydrocarbons and oxygenated compounds. Then, the authors used different machine-learning algorithms for model development, averaging their predictions and thus, obtaining a consensus model [193]. A similar approach was later used for the prediction of density and kinematic density of hydrocarbons [175]. However, as noted by the authors, the use of these models was limited to pure compounds, since mixtures presented significant deviations due to blending effects. A recent study by Creton et al. [194] focused on hydrocarbons and jet fuel, specifically examining fuel sorption; mass gain and sorption kinetics, into polymers. In this work, the authors used 33 Functional Group Count Descriptors to describe the compounds in the dataset. Subsequently, PCA was applied to visualize the chemical space, followed by the formulation of mixtures to populate regions with low sample density.

Another group that has made substantial contributions to modeling hydrocarbon and fuel properties with QSPR is based at the University of Buckingham. The researchers developed the Group Contribution Method of the University of Birmingham (GCM-UOB), a functional group classification system. Its first version included 22 functional group identifiers, position and reactivity descriptors. The system was initially used to formulate jet fuel, diesel, biodiesel and gasoline surrogates by matching the quantities of surrogate functional groups to that of the target fuels through linear regression [195]. The same year, the system was updated to its second version, GCM-UOB 2.0, which incorporated 10 more molecular descriptors, allowing to identify oxygenated compounds and different types of substitution in aromatic rings. Said version was used to predict fuel ignition quality properties (cetane number, research and motor octane number) from medium sized datasets, containing between 400-700 neat compounds and mixtures. In their work the authors found that models that included mixtures in their training data outperformed models that were trained solely on pure compounds, putting in evidence

the non-additivity of the target properties [196]. Finally, the last revision of this system, GCM-UOB 3.0, incorporated 42 descriptors, accounting for structural features, functional group interaction and fuel reactivity. GCM-UOB 3.0 has been successfully used for predicting plethora of properties, including the cetane number, research and motor octane number, surface tension, liquid density, yield sooting index, ignition temperature, vapor pressure, among others [176, 188, 197].

The descriptors developed in the previously mentioned studies may serve as a starting point for the study of the oxidation stability with QSPR. However, it may be necessary to develop new descriptors that encode molecular features specifically related to the oxidation phenomenon.

3.2.1.2 Mixture representation of fuels

Traditional QSPR was originally developed for its use on pure compounds, nevertheless, real world applications of chemistry need the consideration of mixtures. Mixtures have posed a challenge for QSPR, early efforts included using a set of descriptors for each component in the mixture and weight their contribution based on the molar ratio through a mixing rule, the main disadvantages being that the number of descriptors rapidly escalated according to the number of mixture components and the difficulty to obtain an adequate mixing rule [198–200]. Said limitations were overcome by the centroid approximation, where the descriptors of a "pseudomolecule", X_{mix}, are determined by calculating the mean mole weighted average molecular descriptors:

$$X_{\text{mix}} = \sum_{i=1}^{N} x_i \cdot X_i, \tag{3.2}$$

where N is the number of compounds in the mixture, and x_i and X_i are the mole fraction and the molecular descriptor value of the i-th compound.

Even though this method has no mechanistic basis and doesn't consider interactions, it provides good predictions, keeps the number of descriptors constant and utilizes the same computational machinery as traditional QSPR [198, 201].

In the case of complex mixtures, such as fuels, GCxGC has been used to identify their chemical composition. However, this analytical technique only provides information about the hydrocarbon families and carbon number of the compounds, without identification of individual chemical species. Thus, QSPR studies that use information from GCxGC data have relied on several strategies to represent these complex matrices:

- **Selecting one representative molecule:** In this approach, a single molecule represents each bin. However, this is often an oversimplification, as it does not account for isomers

or branching effects [200]. Some efforts to overcome such limitation involve the use of the Modified Weighted Average (MWA) method, which was used to correlate a composition matrix and a property matrix to the target property [202]. While this method improved the model's predictions, it added the complexity of requiring additional experimental measurements and parameters.

- **Sampling from a selection of possible representative molecules:** Isomers are considered in this approach, but the sampling method can be computationally expensive [200].

- **Formulating a surrogate mixture:** While it allows to mimic the properties of fuel with a limited amount of compounds, it requires additional property measurements and may have the same disadvantages as the "selecting one representative molecule" approach.

Hall et al. [200] proposed an approach to address the previously mentioned limitations: the use of Probabilistic Mean Quantitative Structure-Property Relationship (M-QSPR). In this method, a software is used to generate isomers for each GCxGC bin. Molecular descriptors are then calculated and averaged for each bin, with the resulting values scaled by the mole fraction of the corresponding bin. These scaled values are summed to obtain the M-QSPR representation of the samples. This approach enables the development of a model applicable to both pure compounds and real fuels.

In this thesis, we focused on the development of vector descriptors for the modeling of our target property: the Induction Period. As previously mentioned, this type of descriptors have been already widely used for the study of different properties of hydrocarbons. Furthermore, vector descriptors can be easily generalized to account for complex mixtures, by following frameworks such as M-QSPR.

3.2.2 Machine-learning

Modeling algorithms act as the machine that transforms inputs, such as molecular descriptors, into outputs. While simple relationships can be captured using basic algorithms like linear regression, more complex problems involving intricate data and relationships require the use of more advanced techniques, for example,Machine-Learning (ML) algorithms.

Machine-Learning is a branch of artificial intelligence that focuses on developing systems capable of learning and making predictions or decisions without being explicitly programmed. Instead, ML algorithms find patterns in data to build models that generalize to new and unseen data [203]. ML has become a key tool in various scientific and industrial fields, including QSPR modeling, where it aids in predicting the properties of chemical compounds based on their molecular descriptors.

3.2.2.1 Machine-learning paradigms

Machine-learning paradigms are the foundational approaches that guide how algorithms learn from data and make predictions or decisions. These paradigms define the structure of learning based on the availability and nature of data, as well as the desired outcomes. Thus, there are three paradigms [204]:

- **Supervised Learning:** In supervised learning, the model is trained on labeled data, where the input-output mapping is known. This category includes:
  - **Regression:** This task involves training a model to obtain the relationships between an input and a continuous output [205]. In QSPR modeling, regression algorithms are often used to predict properties that continuous properties, such as boiling points, density, or vapor pressure [188, 206].
  - **Classification:** Classification models are used to predict discrete categories. For instance, in QSPR, classification can be applied to predict binary outcomes, such as whether a compound is a drug or a nondrug, or whether a compound exhibits a specific activity [207].
- **Unsupervised Learning:** This type of learning involves data without explicit labels, aiming to uncover hidden patterns or structures. Techniques like clustering (e.g., *k*-means or hierarchical clustering) are useful for grouping samples based on their similarity [208]. Dimensionality reduction techniques, such as PCA [209] and t-distributed Stochastic Neighbor Embedding (t-SNE) [210], are also employed to visualize high-dimensional spaces effectively.
- **Reinforcement Learning:** In reinforcement learning, models learn optimal actions by interacting with an environment and receiving feedback through rewards or penalties. In chemistry, this paradigm has been used for molecule generation, geometry optimization, retrosynthetic pathway search, among other applications [211].

While these paradigms are useful for classifying model learning frameworks, more recent research involves blends across said categories. For instance, semi-supervised learning. This combines aspects of supervised and unsupervised learning, utilizing a small labeled dataset and a larger unlabeled dataset. This approach can be beneficial in QSPR scenarios where obtaining labeled data (e.g., experimental property measurements) is expensive or time-consuming [204].

3.2.2.2 Machine-learning algorithms

Having discussed the paradigms of machine learning, it is important to explore the various algorithms that bring these paradigms to life. Machine-learning algorithms are the tools that

enable systems to process data and derive meaningful insights, tailored to specific tasks such as regression or classification.

- **Linear Models:** Ridge and Least Absolute Shrinkage and Selection Operator (LASSO) regression are extensions of traditional linear regression that allow the analysis of high dimensional data. These methods include regularization terms that can address multicollinearity and redundancy in descriptors, while retaining easy interpretability. These models are particularly useful when the relationships between descriptors and properties are approximately linear [212].
- **Support Vector Machines (SVMs):** These algorithms are useful for regression and classification tasks. They find the optimal decision boundary that separates different classes of data. It can be combined with the use of non-linear kernels, which map the data into higher dimensions, allowing to model complex relationships [213].
- **Random Forests and Gradient Boosting Machines:** A family of methods based on decision trees. A decision tree is a hierarchical model used for classification and regression tasks. Internal nodes represent features or variables, branches represent decision rules or feature splits, and leaves represent either a class label (for classification) or a continuous value (for regression). The tree recursively splits the data into subsets based on feature-based rules, continuing until the data is assigned to a specific class or predicted value at the leaves [212, 214]. Gradient Boost Machines use successive models consisting of weak learners (e.g. simple decision trees). The boosted models tend to outperform random forests [215].
- **Artificial Neural Networks (ANNs):** These models mimic the behavior of biological neurons, consisting of interconnected nodes (neurons) organized in layers. These networks learn by adjusting the weights of connections through iterative training, enabling them to capture complex relationships between inputs and output [216].
- **Deep Learning Models:** Techniques like convolutional neural networks (CNNs) and graph neural networks (GNNs) are increasingly used to process molecular graphs and extract features directly from structural representations. These models avoid reliance on predefined descriptors by learning representations directly from raw molecular structures [217].

3.2.2.3 Hyper-parameter tuning

The performance of the previously described machine-learning algorithms depends on hyper-parameter selection. A hyper-parameter differs from regular model parameters because it cannot be directly estimated from the data learning process, and must be set before training a ML model since they determine the model architecture [218]

Hyper-parameter tuning involves selecting the set of values for the data set. For instance, consider a dataset consisting of points (x, y), where y is a function of x, plus some noise. In this scenario, fitting a polynomial serves as our "machine-learning" algorithm.

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_0 \tag{3.3}$$

Here, the degree of the polynomial, n, acts as the hyper-parameter of the algorithm. Choosing the appropriate degree is crucial, since a low degree might lead to underfitting, where the polynomial is too simple to capture the underlying data pattern. On the other hand, choosing a high degree might lead to overfitting, where the polynomial captures noise rather than the true trend. When a model is said to be overfitted, it "memorizes" the input data, thus compromising its performance on unseen data (see figure 3.5).

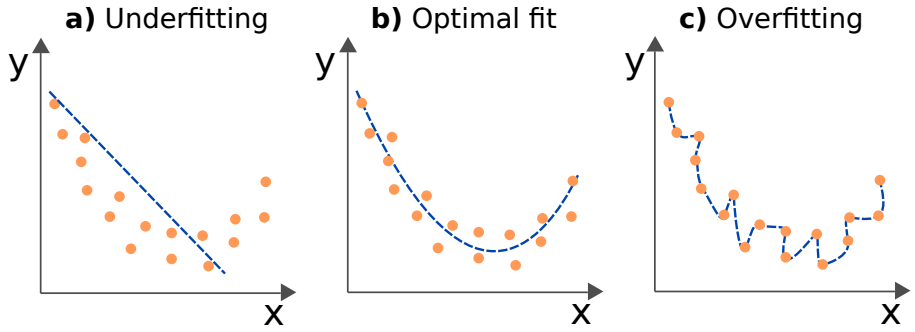


FIGURE 3.5 – Illustration of model fitting with varying polynomial degrees: **a)** Underfitting with a low-degree polynomial fails to capture the complexity of the data. **b)** Optimal fit with an appropriately chosen degree balances bias and variance, accurately capturing the data trend. **c)** Overfitting with a high-degree polynomial fits the noise in the data, resulting in poor generalization.

There are several methodologies to search for the optimal set of hyper-parameters. Two of the simplest methods are random search and grid search. In random search, random values are sampled from the hyper-parameter space, while in grid search, a predefined grid of hyper-parameter values is evaluated, and the model's accuracy is assessed for all possible combinations [219]. An alternative for these traditional methods of hyper-parameter tuning is Bayesian optimization, which aims to find the global optimum of a black-box function by constructing a probabilistic model, and using it to decide which combination of hyper-parameters to next evaluate [220].

The hyper-parameter tuning process depends not only on the method used for hyper-parameter search but also on the validation step. Model validation is essential as it helps evaluate both the optimal set of hyper-parameters and the overall accuracy of the model.

3.2.2.4 Model validation

Regardless of the chosen algorithm, model validation is a crucial step before deploying a model. The validation step has two purposes, to identify the hyper-parameter setting that result in the most realistic prediction performance, and to asses the model's accuracy and generalizability.

Model validation involves an optimization step that requires an objective function to be maximized or minimized. For error functions, also referred to as loss functions, the goal is typically to minimize their value [221]. For example, in regression tasks, the Root Mean Square Error (RMSE) is commonly used to measure model's performance:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \tag{3.4}$$

where:

- y_i is the i-th reference value,
- $\hat{y}_i$, the i-th predicted value,
- and N is the number of samples.

Thus low RMSE values indicate a higher model performance. Traditionally, model validation involves splitting the data into training and test sets. The training set is used for model development and hyper-parameter tuning, while the test set remains untouched during the learning process. It is reserved exclusively for evaluating the model's performance and assessing its accuracy on unseen data [222]. The ratio of training to test data can vary depending on the size of the dataset, with common splits being 70:30 or 80:20. Ideally, the split should cover the full range of the dependent variable y and account for any inherent structures in the data, such as sample origin, treatment conditions, or instrumental replicates [223].

There are different methodologies to perform hyper-parameter tuning, in this work we will mention the most common ones. When dealing with small datasets (e.g., $N < 40$), Leave-One-Out (LOO) cross-validation is often recommended. In LOO, each sample is sequentially excluded from the dataset, and the remaining $N - 1$ samples are used to train the model. The excluded sample is then used for validation. This process is repeated until each sample has been used for validation once. While LOO assesses model performance for each data point, it tends to underestimate the model error, making it less reliable for model selection compared to other methods [223, 224].

k-fold cross-validation is a widely used technique for medium-sized datasets (see figure 3.6). It divides the data into k segments (or folds), with each fold serving as an internal validation set while the remaining $k - 1$ folds are used for training. This process is repeated until each fold has been used as the validation set exactly once. The RMSE, or other appropriate metric, is calculated for each fold, and the results are averaged to obtain an overall estimate of model

performance. This method mitigates the risk of having a particularly easy or difficult validation set, providing a more reliable estimate of the model's generalizability [212].

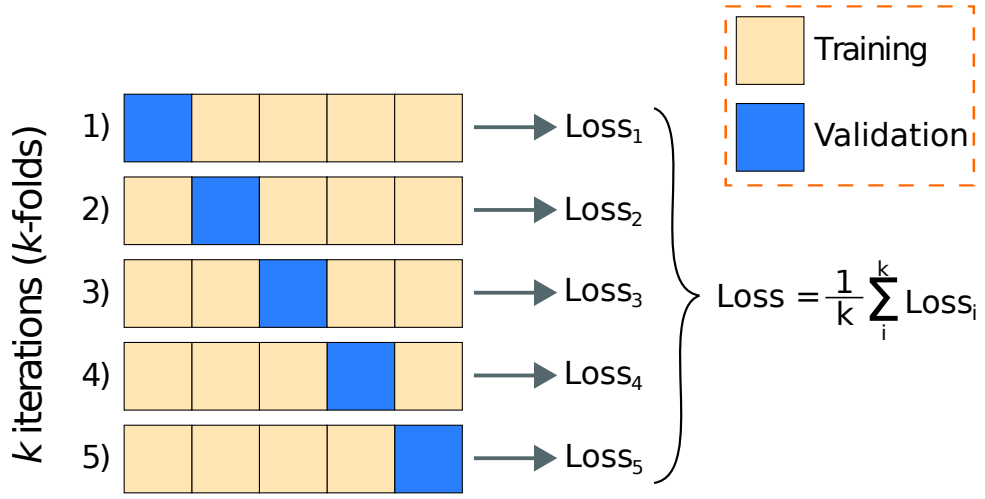


FIGURE 3.6 – Illustration of k-fold cross-validation with $k = 5$. The dataset is divided into 5 equal-sized folds, where each fold (in blue) is used as a validation set exactly once, while the remaining folds (in orange) are used for training. The process is repeated k times, and the model's performance is evaluated using a loss function (e.g., RMSE). The final loss is calculated as the average of the losses obtained from all k iterations.

An alternative to k-fold cross-validation is bootstrap cross-validation, where the data is sampled with replacement. This means that some data points may appear in multiple training sets, while others may be excluded entirely. Similar to k-fold cross-validation, the model is trained and evaluated multiple times, but the sampling process allows for a more varied set of training and validation data. Bootstrap cross-validation is particularly useful when dealing with smaller datasets and helps to provide more robust estimates of model performance, such as RMSE, compared to LOO validation [225].

The hyper-parameter tuning step involves applying one of the search algorithms described in section 3.2.2.3 in combination with a cross-validation method to identify the set of hyper-parameters that minimize the loss function. Once the optimal hyper-parameters are selected, the model is retrained on the entire training set using these parameters. Finally, the trained model is evaluated on the test set to assess its accuracy on unseen data [212].

Modern approaches to model validation involve splitting the data into multiple train and test sets. These methodologies have proven to provide find more optimal hyper-parameters and a better estimation of the model accuracy [222].

3.3 Methodology

3.3.1 Molecular descriptors and model features

As noted in section 3.2.1.2, we chose to focus on descriptor vector representations, specifically, Functional Group Count Descriptors (FGCD). This decision stems from the demonstrated success of this approach in modeling other hydrocarbon properties, and its potential to be easily extended to account for hydrocarbon mixtures. Thus, as a starting point, we based our work on the descriptors developed by [194], who introduced 33 descriptors for predicting fuel sorption into polymers. The FGCDs listed in table 3.2, primarily focus on simple molecular fragments, such as primary, secondary, tertiary and quaternary carbons, number of aromatic carbons, etc.

Table 3.2 – List of the model features and their descriptions, used by Creton et al. [194].

Molecular descriptor	Description
[H]	Number of H atoms.
[C,c]	Number of C atoms.
[CX4H3]	Aliphatic C with 1 further total connections, with 3 further hydrogen.
[CX4H2]	Aliphatic C with 2 further total connections, with 2 further hydrogen.
[CX4H1]	Aliphatic C with 3 further total connections, with 1 further hydrogen.
[CX4H0]	Aliphatic C with 4 further total connections, with 0 further hydrogen.
[CX3H1]	Aliphatic C with 2 further total connections, with 1 further hydrogen.
[CX4H2R]	Aliphatic C with 2 further total connections, with 2 further hydrogen, in a ring.
[CX4H1R]	Aliphatic C with 3 further total connections, with 1 further hydrogen, in a ring.
[cX3H1](:*):*	Aromatic C with 0 further total connections, with 1 further hydrogen.
[cX3H0](:*)(:*)*	Aromatic C with 0 further total connections, with 0 further hydrogen.
[cX3H0](:*)(:*):*	Aromatic C with 0 further total connections, with 0 further hydrogen.
[cX3H0]-[cX3]	Aromatic C with 2 further total connections, with 0 further hydrogen.
[cX3H0](:*)(:*)(-[CX4H2R])	Aromatic C with 0 further total connections, with 0 further hydrogen bound to aliphatic C with 1 further total connections, with 2 furthed hydrogen, in a ring.
[CX4H2]-[CX4H1]-[CX4H2]	Aliphatic C with 1 further total connections, with 2 further hydrogen, bound to aliphatic C with 1 further total connections, with 1 further hydrogen, bound to aliphatic C with 1 further total connections, with 2 further hydrogen.
[C][C]([!CX1])([!CX1])[!CX1]	2 bound aliphatic C with 0 further total connections, with 3 non C atoms.
[!C][C]([C])([C])[C]	Aliphatic C with 3 further aliphatic C and 1 further non C atom.
[C][CR]([!C])([!C])[C]	Aliphatic C in a ring connected to 2 aliphatic C and 2 non C atoms.
[C][CR]([!C])([C])[C]	Aliphatic C in a ring connected to 3 aliphatic C and 1 non C atom.
[C]=[C]([C])[!C]	Aliphatic C connected to 1 aliphatic C with a double bond, to 1 aliphatic C with single bond, and a non C atom.
[CX3H1]=[CX3H1]	Bound of two aliphatic C with 1 further total connection, with 1 hydrogen.

Continued on next page

TABLE 3.2 – continued from previous page.

Molecular descriptor	Description
[c][CX4H3]	Aliphatic C with 1 further aromatic C and 3 further hydrogen.
[c][CX4H2]	Aliphatic C with 1 further total connection, with 1 further aromatic C and 2 further hydrogen.
[c][CX4H1]	Aliphatic C with 2 further total connections, with 1 further aromatic C and 1 further hydrogen.
[R]	Number of atoms in a ring.
aromatic_rings	Number of aromatic rings.
non-aromatic_rings	Number of non-aromatic rings.
aliphatic_rings	Number of aliphatic rings.
number_of_rings	Total number of rings.
MM	Molecular mass.
[C;R]	Aliphatic C atom, in a ring.
[c;R]	Aromatic C atom, in ring.
C1CCCCC1	Six C atom aliphatic ring.

However, certain shortcomings in this set of descriptors became apparent. One key issue was degeneracy, when the descriptor matches substructures representing different molecular features (see figure 3.7). Furthermore, the reported SMARTS patterns did not account for molecular features relevant for hydrocarbon stability, such as ortho-, meta-, and para- positions in aromatic compounds, or the number of allylic carbons in olefins and conjugated olefins.

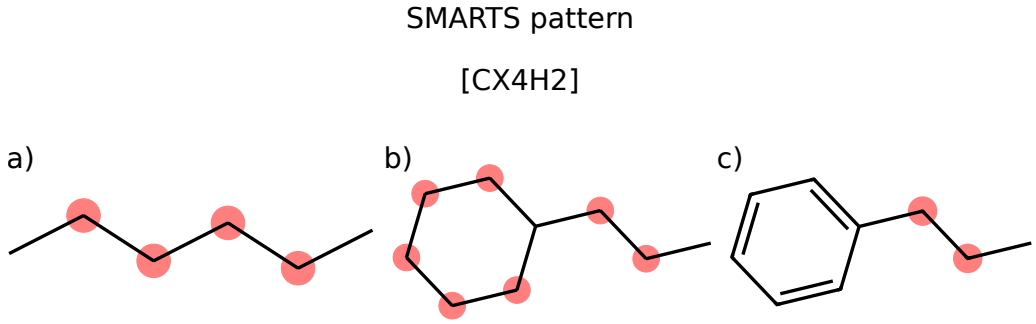


FIGURE 3.7 – Degeneracy for the SMARTS pattern [CX4H2] for **a)** *n*-hexane, **b)** *n*-propylcyclohexane, and **c)** *n*-propylbenzene. The molecular descriptor cannot differentiate between CH_2 groups in a chain, an aliphatic ring, or at benzylic sites.

The literature and our results presented in section 2.5, show that molecular fragments have different impact on the oxidation stability according to their surrounding, e.g., whether a carbon is present in an aliphatic chain, an allylic or a benzylic site [16, 50, 89, 91, 226]. Following this reasoning, we expanded the set of descriptors proposed by Creton et al. [194] to account for the reactivity of the various hydrocarbon families in our database. To achieve this, we included specific descriptors that reduce degeneracy by considering substitution positions in aromatic rings, the number of allylic carbons for non-conjugated and conjugated double bonds, unfused aromatic rings, and other relevant features. The calculation of molecular descriptors

was performed using the SMARTS-matching functionalities from the RDKit toolkit [227]. The final set of SMARTS patters is presented in table 3.3, labeled from X1 to X42. For instance, the FGCD labeled X3 denotes the number of CH_3 groups bonded to a carbon atom in a ring. The molar mass (MM) of the pure compounds was also computed and used as a descriptor (labeled X16). Furthermore, the temperature (in K) at which the experiments were conducted was used as an additional model feature (labeled T). The complete database containing molecule names, SMILES codes and the experimental IP values at the corresponding temperatures, is available in Appendix B.

Table 3.3 – List of model features and their descriptions used in this study.

Label	Model feature	Description
T	Experiment temperature	Experiment temperature.
X1	[H]	Number of hydrogen atoms.
X2	[$([CX4H3][C!R])]	Aliphatic C with 1 further connection, with 3 further H atoms, not attached to a C atom in a ring.
X3	[$([CX4H3][CR])]	Aliphatic C with 1 further connection, with 3 further H atoms, attached to a C atom in a ring.
X4	[$([CX4H2]([C!R])[C!R])]	Aliphatic C with 2 further connections, with 2 further H atoms, attached to two C atoms not in a ring.
X5	[$([CX4H2!R][CR])]	Aliphatic C with 2 further connections, with 2 further H atoms, not in a ring, attached to a C atom in a ring.
X6	[$([CX4H1]([C!R])([C!R])[C!R])]	Aliphatic C with 3 further connections, with 1 further H atom, not attached to C atoms in a ring.
X7	[$([CX4H0]([C!R])([C!R])([C!R])[C!R])]	Aliphatic C with 4 further connections, with 0 further H atoms, not attached to a C atom in a ring.
X8	[$([CX4H0!R][CR])]	Aliphatic C with with 4 further connections, with 0 futher H atoms, not in a ring, attached to a C atom in a ring.
X9	[CX4H1R]	Aliphatic C with 3 further connections, with 1 further H atom, in a ring.
X10	[CX4H2!R] [cR1]1[cR1][cR1][cR1][cR1][cR1]1	Aliphatic C with 2 further connections, with 2 further H atoms, not in a ring, attached to an unfused aromatic ring.
X11	[CX4H1!R] [cR1]1[cR1][cR1][cR1][cR1][cR1]1	Aliphatic C with 3 further connections, with 1 further H atoms, not in a ring, attached to an unfused aromatic ring.
X12	[CX4H2!R][CX4H1!R] [cR1]1[cR1][cR1][cR1][cR1][cR1]1	CH_2—CH_1 attached to an unfused aromatic ring.
X13	[CX4H0!R] [cR1]1[cR1][cR1][cR1][cR1][cR1]1	Aliphatic C with 4 further connections, with 0 further H atoms, not in a ring, attached to an unfused aromatic ring.
X14	MM	Molecular mass.
X15	[c;R]	Aromatic C in a ring.
X16	C1CCCCC1	Six carbon atom aliphatic ring.
X17	[$([CX4!H0][CX3]=[CX3][CX3]), $([CX4!H0][CX3][CX3]=[CX3])]	Aliphatic C with 4 further connections, with at least 1 further H atom, attached to a carbon in a conjugated double bond.
X18	[$([CX4!H0][CX3]=[CX3]); !$([CX4!H0][CX3][CX3]); !$([CX4!H0][CX3]=[CX3][CX3])]	Aliphatic C with 4 further connections, with at least 1 further H atoms, attached to carbon with a double bond.
X19	[$([CX4H0][CX3]=[CX3]); !$([CX4!H0][CX3][CX3]); !$([CX4!H0][CX3]=[CX3][CX3])]	Aliphatic C with 4 further connections, with 0 further H atoms, attached to carbon with a double bond.
X20	[$([CX4!H0]([CX3])[cccccc])]	Aliphatic C with 4 further connections, with at least 1 further H atom, attached to an aromatic ring and a carbon with a double bond.
X21	[CX4]ccc[CX4]	Aliphatic C atoms attached to an aromatic ring, in meta- position.
X22	[CX4]cc[CX4]	Aliphatic C atoms attached to an aromatic ring, in ortho- position.
X23	[CX4]cccc[CX4]	Aliphatic C atoms attached to an aromatic ring, in para- position.
X24	[$([CX4!H0!R]([cccccc])[cccccc])]	Aliphatic C with 4 further connections, with at least 1 further H atom, not in a ring, attached to two aromatic rings.

Continued on next page

Table 3.3 – continued from previous page.

Label	Model feature	Description
X25	[$([CX4!H0R]([cccccc])[cccccc])]	Aliphatic C with 4 further connections, with at least 1 further 1 atom, in a ring, attached to two aromatic rings.
X26	[CX4H3] [cR1]1[cR1][cR1][cR1][cR1][cR1]1	Aliphatic C with 1 further connections, with 3 further H atoms, attached to an unfused aromatic ring.
X27	[c][CX4H2r6]	Aliphatic C with 2 further connections, with 2 further H atoms, in a 6-member ring, attached to an aromatic C atom.
X28	[c][CX4H2r5]	Aliphatic C with 2 further connections, with 2 further H atoms, in a 5-member ring, attached to an aromatic C atom.
X29	c12ccccc1cccc2	Naphthalene ring.
X30	[$([CX4H3][c]); !$([CX4H3] [cR1]1[cR1][cR1][cR1][cR1][cR1]1)]	Aliphatic C with 1 further connection, with 3 further H atoms, attached to a partially hydrogenated naphthalene ring.
X31	[$([CX4H2R][c]);!$([CX4H2] [cR1]1[cR1][cR1][cR1][cR1][cR1]1)]	Aliphatic C with 2 further connections, with 2 further H atoms, in a ring, attached to a partially hydrogenated naphthalene ring.
X32	[$([CX4H1R][c]);!$([CX4H1] [cR1]1[cR1][cR1][cR1][cR1][cR1]1)]	Aliphatic C with 3 further connections, with 1 further H atoms, in a ring, attached to a partially hydrogenated naphthalene ring.
X33	[cR1]1[cR1][cR1][cR1][cR1][cR1]1	Unfused aromatic ring.
X34	[$([CR]([!c])[!c])]	Aliphatic C in a ring, not attached to aromatic carbons.
X35	[$([CX4H2R] [cR2]1[cR2][cR1][cR1][cR1][cR1]1)]	Aliphatic C with 2 further connections, with 2 further H atoms, in a ring, attached to an aromatic ring.
X36	[$([CX4H1R] [cR2]1[cR2][cR1][cR1][cR1][cR1]1)]	Aliphatic C with 3 further connections, with 1 further H atom, in a ring, attached to an aromatic ring.
X37	[CX4H3]ccc[CX4H3]	Methyl groups attached to an aromatic ring in meta- position.
X38	[CX4H3]cc[CX4H3]	Methyl groups attached to an aromatic ring in ortho- position.
X39	[CX4H3]cccc[CX4H3]	Methyl groups attached to an aromatic ring in para- position.
X40	[$([CX4H3][c]); !$([CX4H3] [cR1]1[cR1][cR1][cR1][cR1][cR1]1); !$([CX4H3][c][c][C])]	Aliphatic C with 1 further connection, with 3 further connections, attached to a naphthalene ring.
X41	[$([CX4H2!R][c]); !$([CX4H2] [cR1]1[cR1][cR1][cR1][cR1][cR1]1); !$([CX4H2][c][c][C])]	Aliphatic C with 2 further connection, with 2 further connections, not in a ring, attached to a naphthalene ring.
X42	[$([CX4H2R][c]); !$([CX4H2] [cR1]1[cR1][cR1][cR1][cR1][cR1]1); !$([CX4H2][c][c][C])]	Aliphatic C with 2 further connection, with 2 further connections, in a ring, attached to a naphthalene ring.

3.3.2 Modeling method

The development of QSPR-based models using ML approaches has been widely used for property prediction. These approaches highlight hidden correlations between a property and some features encoding information about the sample. Reviews already exist containing detailed elements regarding best practices in developing such models [189, 228]. The employed methodology is summarized hereafter.

3.3.2.1 Data sets

The accuracy of ML models heavily relies on the quality of the reference data, making the database employed in model development one of the most crucial elements for such studies. The database used in this work is the one presented in section 2.4.1, which contains 220 reference IP values from 95 compounds, obtained by following the experimental procedure previously described in section 2.4.2. As shown in figure 3.8**a**, the IP distribution is positively skewed, with approximately 80% of the measurements presenting an IP lower than 10 h. The severe skewness in our data may affect the accuracy of the models. Therefore, a preprocessing step, such as log-transforming the data may be needed (see figure 3.8**b**).

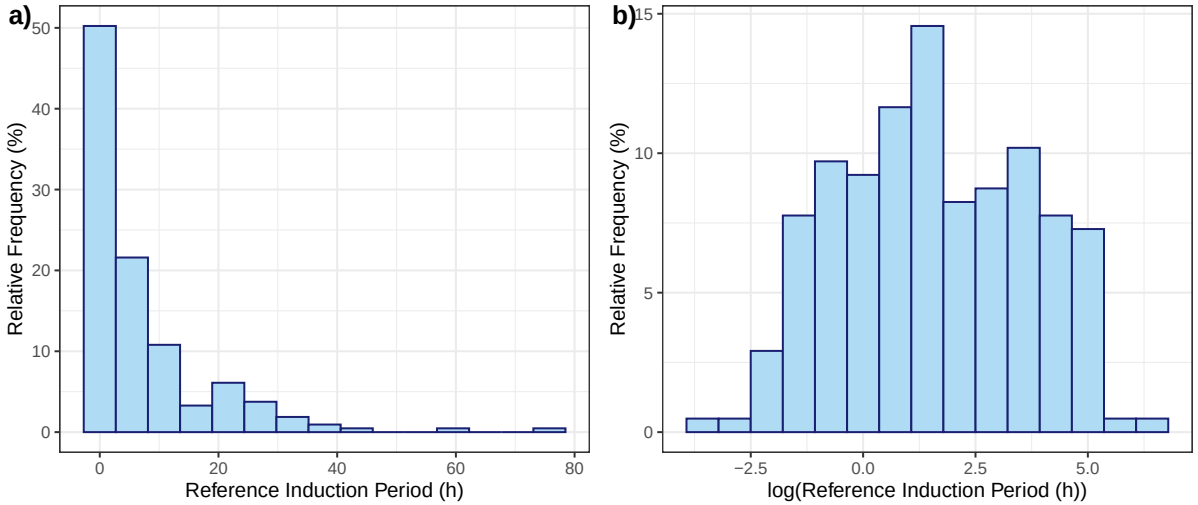


FIGURE 3.8 – Induction period distribution for the compounds in the database in **a)** hours and **b)** log(hours).

3.3.2.2 Machine learning algorithms

We employed two ML algorithms for model development: Support Vector Machine (SVM) [229], and eXtreme Gradient Boosting (XGBoost), reported as one of the most efficient methods for classification and regression [230]. These algorithms were chosen for their capability to handle highly non-linear data, as the Induction Period exhibits a complex, non-linear relationship with molecular features such as paraffinic chain length and the degree of substitution in olefins. Modeling was performed using the R programming language [231], along the `caret` [232], `kernlab` [233] and `xgboost` [234] packages for model training and implementation of the SVM and XGBoost algorithms, respectively.

Given our previously defined data set, $\mathbf{X}$ with dimensions $N \times M$, where N is the number of samples (221) and M, the number of descriptors (42), and an output vector $\mathbf{y}$ of length N, the SVM finds a regression function (SVR, Support Vector Regression), $f(\mathbf{x})$, through the equation:

$$f(\mathbf{x}) = \sum_{i,j=1}^{N} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}_j) + b \tag{3.5}$$

where:

- α_i and α_i^* are the Lagrange multipliers that satisfy the condition $0 \leq \alpha_i, \alpha_i^* \leq C$,
- C, the cost hyper-parameter,
- b a bias-term, and
- $K(\mathbf{x}_i, \mathbf{x}_j)$ the kernel function.

Non-linear kernel functions were used to map data into high dimensions, allowing to find a hyperplane that linearly separates the data. In this work, we used the Radial Basis Function (RBF) kernel [235], which has the form:

$$K_{\text{RBF}}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}\right) \tag{3.6}$$

where σ is a parameter related to the decay of the exponential function.

On the other hand, XGBoost uses weak learners, such as linear models and shallow regression trees, to train a model in an additive manner. In this work, we will hereafter refer to these two implementations as XGB Linear and XGB Tree. Thus, the model is built by minimizing the following objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x})\right) + \gamma T_t + \frac{1}{2}\lambda||\omega_t||^2 \tag{3.7}$$

where:

- $\mathcal{L}^{(t)}$ is the objective function at the iteration t,
- $l\left(y_i, \hat{y}^{(t-1)}\right)$ is the loss function evaluated at the i-th reference value y_i and the predicted value at the $(t-1)$ iteration, $\hat{y}_i$ plus, a tree structure $f_t(\mathbf{x})$,
- γ is a hyper-parameter that penalizes the complexity of the tree based on the number of leaves, and
- T, and λ, regularization hyper-parameters that smooth the leaf weights ω to avoid overfitting.

3.3.2.3 Model development

In this work, we applied a nested Cross-Validation (CV) procedure to find the optimal hyper-parameters and estimate the model performance (see figure 3.9). The first round consisted in

splitting the data into k external folds. Each fold was used once as a test set, while the samples in the remaining k-1 folds were used as a training set. Then, the training set was further split into i internal folds, each fold was left-out as a validation set, while surrogate models were trained on the remaining $i - 1$ folds. Thus, the external validation step was used for assessing model accuracy, while the inner validation step was used for hyper-parameter tuning by using a grid search method.

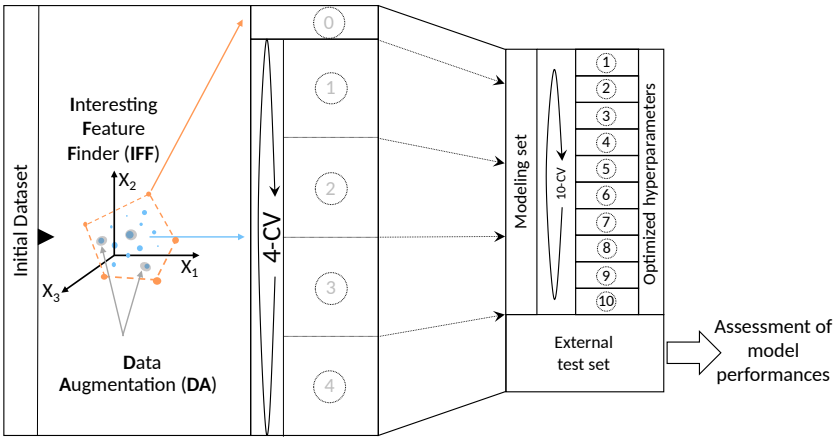


FIGURE 3.9 – Model validation workflow used in this work. An individual QSPR model is built using ML methods (SVR or XGBoost) and molecular descriptors within both internal (10-CV) and external (4-CV) cross-validation procedures, followed by its validation on the external test set.

In our study, we used 10 internal folds ($i = 10$) for hyper-parameter optimization and 4 external folds ($k = 4$) for model evaluation. The choice of the internal folds was guided by the small dataset size, ensuring that a maximum number of samples were included in the training set during hyper-parameter tuning. For the external validation, $k = 4$ was selected for two reasons: (1) the training set was already substantial due to the inclusion of unique, fixed samples, and (2) time constraints associated with the computational cost of nested CV.

Database splitting was randomly performed by applying a compound-out strategy [236]. This means that when a molecule was selected for a fold assignment, all its instances (IP measurements at different temperatures) were included as well. Furthermore, to avoid Applicability Domain (AD) violations during model validation, we employed a maximum dissimilarity sampling algorithm.

Our initial attempt involved using the MaxMin Maximum-Dissimilarity algorithm, commonly referred to as the Kennard-Stone algorithm [237]. Broadly speaking, this algorithm selects a subset of k samples from a total set of n samples. The process begins with an initial subset of samples, and at each iteration, the algorithm adds the most dissimilar samples relative to those already selected. A detailed explanation of this algorithm is provided in appendix C. However, the Kennard-Stone algorithm consistently failed to select hydrocarbons from the n-paraffin class, leading to violations of the Applicability Domain.

For this reason, we opted to use the Interesting Feature Finder (IFF) algorithm [238]. This method generates a large number of random vectors, each with a length matching the number of columns in the dataset **X**. These vectors are projected onto the rows of the **X** matrix to identify samples with the smallest and largest dot products. Samples appearing frequently among those with extreme dot products are considered the most dissimilar in the dataset. In this work, we first applied auto-scaling, which involves mean-centering and scaling each column of the matrix by dividing it by its standard deviation. Then we projected 10 000 randomly generated vectors on the data matrix. Next, we identified the compounds selected by the algorithm as extreme samples, at least 1% of the times. This method identified 25 dissimilar compounds, such as cyclopentane, allylbenzene, and *n*-eicosane, which were fixed in the training set, as a "Fold 0" [236, 239, 240].

Since we are dealing with a regression task, we used Root Mean Square Error (RMSE) as the loss function, as defined in equation (3.4). RMSE is denoted as RMSEC, RMSECV, and RMSEP when applied to the error of the Calibration (training), Cross-Validation, and Prediction (test), respectively.

3.3.2.4 Data Augmentation (DA)

As previously shown in table 2.1, our database presents an imbalance in the hydrocarbon families. For instance, mono-aromatics, paraffins, mono-naphthenes and mono-olefins represent roughly 80% of the compounds in the database, while the remaining 9 hydrocarbon families correspond to 20% of the database. This results in some molecular descriptors used for molecular representation being unique to a specific sample or hydrocarbon class. Consequently, if such a sample is included in the test set, the model will fail to predict its IP value accurately. Thus, the use of synthetic data, also known as data augmentation, may help to improve the performance of the models [241].

In our work, we generated synthetic data using measured samples as reference, applying slight structural modifications that, in principle, should not affect their reactivity. This process perturbed the input matrix **X**. Subsequently, we introduced noise, σ_{noise}, to the IP of the "reference molecule" used to generate the synthetic sample. The magnitude of the noise was chosen based on the method's repeatability (±11.7%). For example, we found that at 120 °C, allylbenzene is approximately 90 times more reactive than toluene due to the formation of a resonance-stabilized free radical following H-abstraction at the carbon between the double bond and the aromatic ring. Based on this observation, we assumed that adding a methyl group to the aromatic ring of allylbenzene would not significantly affect its overall reactivity. Consequently, *p*-allyltoluene, which results from adding a methyl group to the para- position of the aromatic ring in allylbenzene, is expected to have an IP similar to that of allylbenzene (see figure 3.10).

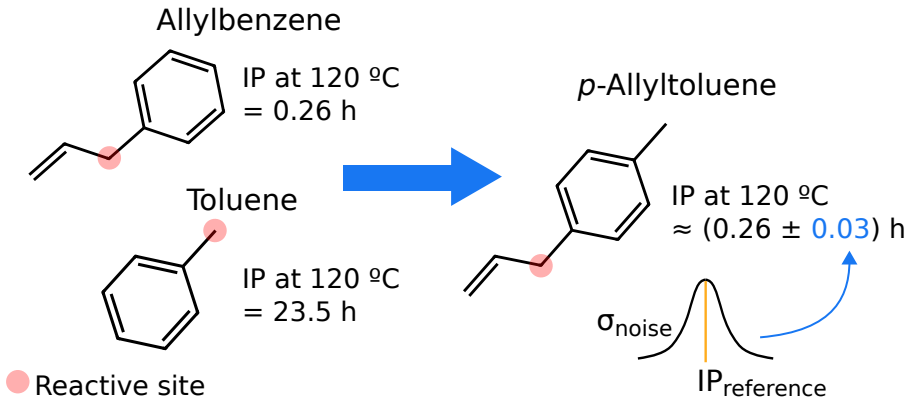


FIGURE 3.10 – Schematic of the Data Augmentation strategy followed. The selected under-represented molecules were perturbed by adding molecule features with negligible impact on their reactivity, while their corresponding Induction Period, $IP_{reference}$ values were multiplied by a random number proportional to the relative method repeatability, σ_{noise}.

In total, 36 molecules were added with this methodology. The additional molecules were only included in the Fold-0, thus the test sets consisted exclusively of experimentally analyzed molecules. The complete list of augmented structures, along with their corresponding hypothetical IP values, can be found in Appendix D.

3.3.2.5 Model explanation

Machine learning algorithms can identify relationships in complex and non-linear data. However, the interpretability of certain algorithms, such as SVM-RBF, is often limited. To address this, the field of Explainable Machine Learning (XML) emerged, focusing on the need to understand predictions generated by machine learning models. The primary goal of XML is to achieve "explainability," making specific aspects of a system comprehensible to humans [242].

Numerous methods for model explanations exist. In this work, we focus on the widely known model-agnostic method called SHapley Additive exPlanations (SHAP) [243]. SHAP is a game-theory-based approach that utilizes the Shapley value framework, which was originally developed to assign payouts to players based on their contributions to the total payout [244]. The formula for calculating Shapley values is:

$$\phi_i(f,x) = \phi_i = \sum_{z' \subseteq x' \setminus \{i\}} \frac{|z'|!\,(M-|z'|-1)!}{M!} \left[f_x\left(z' \cup \{i\}\right) - f_x\left(z'\right) \right] \quad \text{for } i = 1,\ldots,M \tag{3.8}$$

where:

- $\phi_i(f,x)$ is the Shapley value for the model feature i, a function of the predictive model f, and a specific instace or row of the input data, x,

- z', a subset of features from x that doesn't include i,
- $x' \setminus \{i\}$, the set of all features in x except for feature i,
- $|z'|$, the number of features in the subset z',
- M, the total number of features in x,
- $f_x(z' \cup \{i\})$, the output of the model when only the features in z' are considered, and
- $f_x(z')$, the output of the model when the features in z' and feature i are included.

In equation (3.8), the term $[f_x(z' \cup \{i\}) - f_x(z')]$ represents the contribution of the feature i to the model output, also referred to as the marginal value. Meanwhile, the term $\frac{|z'|!(M-|z'|-1)!}{M!}$ serves as a weight, scaling the marginal value based on the number of features in the subset z'.

However, there are two main limitations for the accurate calculation of Shapley values. The first is the vast number of possible subsets in multivariate data commonly used in machine learning, given by the formula 2^M. The second is that the original Shapley value implementation assumes the independence of model features, which does not always hold true [245]. To address this, SHAP offers a framework to approximate the exact Shapley values. This is achieved through model-specific approximations, such as Kernel SHAP [243] for non-linear models, or Tree SHAP for tree-based models [246].

SHAP values are widely used because they offer both local and global explanations for model predictions. For example, features with positive SHAP values contribute positively to the prediction, while those with negative values result in a decrease of the predicted value. Moreover, the mean absolute SHAP value reflects the magnitude of each feature's influence. In our work, we used the `shapr` R package, which implements the calculation of the SHAP values as described by Aas et al. [245].

3.4 Results and discussion

3.4.1 Model performance

The entire dataset of experimental IP values was used to derive predictive models. Nested 4-CV and 10-CV were applied as illustrated in figure 3.9, resulting in (i) the splitting of the database into 4 folds for external validation, plus one additional fold containing compounds fixed in the training set, to avoid violation of the applicability domain ; (ii) the splitting of the training set into 10 folds for internal validation and hyper-parameters optimization. To reduce class imbalance effects over hydrocarbon family representation, a data augmentation technique was applied, and new virtual structure/data were used to supplement the Fold-0. Additionally, the effect of applying or not a log-transformation of IP values prior to model development was investigated. Three types of models, involving SVM and XGBoost ML algorithms, were built: SVR with RBF kernel, XGBoost with linear learners, and XGBoost with tree ensemble learners.

All models were developed and validated according to the workflow presented in figure 3.9. Hereafter, we discuss the predictive performance of the obtained ML-QSPR based models for which obtained metrics values are reported in table 3.4 and table 3.5, for models built without and with Data Augmentation, respectively.

As shown in table 3.4, XGBoost Tree outperformed both SVR-RBF and XGBoost Linear, with an RMSEP of 7.00 h (127%). However, the parity plots in figure 3.11 reveal that all three models struggle to predict IP values below 2 h, as indicated by the vertical patterns in the plots.

Table 3.4 – Model performance metrics for SVR-RBF, XGBoost with linear learners and XGBoost with regression trees learners without Data Augmentation. RMSE values were averaged across the 4 external cross-validation folds.

Model	**Optimal hyper-parameters**	**Scale**	**RMSEC**	**RMSECV**	**RMSEP**
SVR-RBF	$\sigma = 0.025$ $C = 15.8$ $\epsilon = 0.3$	hour	6.74	7.53	10.16
XGBoost Linear	n_rounds = 250 eta = 0.05 $\lambda = 2$ $\alpha = 4$	hour	3.09	8.54	8.36
XGBoost Tree	n_rounds = 100 eta = 0.05 max_depth = 12 $\gamma = 0.33$ colsample_bytree = 0.66 min_child_weight=5 subsample = 1	hour	3.62	8.13	7.00
SVR-RBF log-transform	$\sigma = 0.02$ $C = 3.98$ $\epsilon = 0.00$	log hour	0.59 5.34	1.46 7.89	1.50 8.01
XGBoost Linear log-transform	n_rounds = 250 eta = 0.05 $\lambda = 2$ $\alpha = 0$	log hour	0.05 0.54	1.57 9.04	1.16 6.75
XGBoost Tree log-transform	n_rounds = 200 eta = 0.3 max_depth = 3 $\gamma = 0$ colsample_bytree = 1 min_child_weight=0.25 subsample = 0.66	log hour	0.40 3.31	1.43 10.10	1.13 8.12

This issue could be related to two factors; the data is positively skewed (see figure 3.8), and the disparity between low and high IP values, spanning two orders of magnitude. Consequently, the models may fail to predict lower IP values due to their minimal impact on the overall RMSE compared to higher IP values.

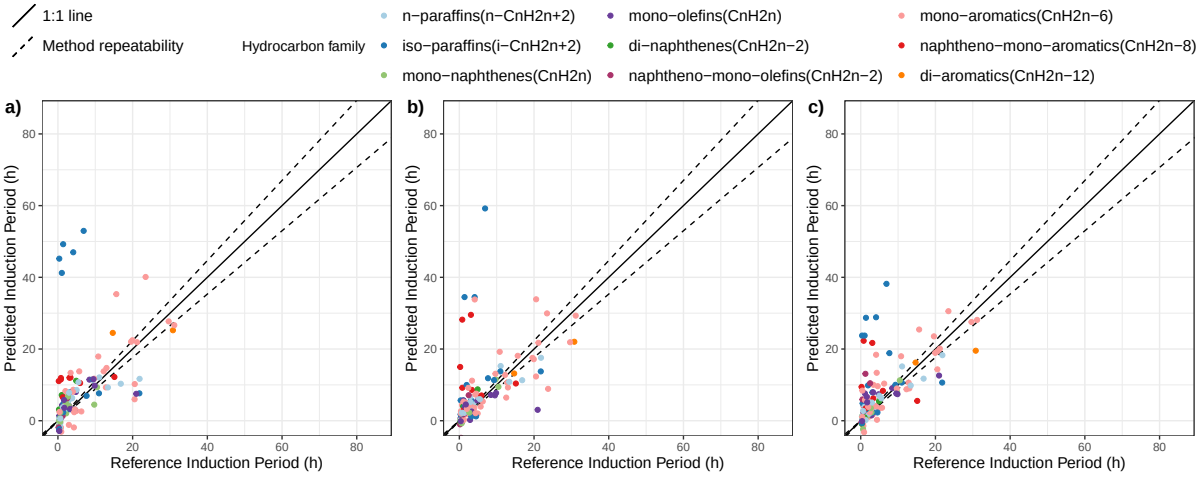


FIGURE 3.11 – Reference vs. predicted IP values for the 4 external CV folds. Shown for **a)** Support Vector Regression, **b)** XGBoost with linear learners, and **c)** XGBoost with regression trees. Without log transformations or Data Augmentation.

On the other hand, models obtained after log-transforming the IP values presented an RMSEP 20% lower than non-transformed models. This could be explained by the reduction of data skewness after applying the log-transform, resulting in all samples contributing more equally to the model error and a better fit (see figure 3.8**b**). For this set of models, XGBoost Linear presented the highest accuracy, with an RMSEP of 6.75 h, corresponding to a relative error of 122.9%. XGBoost Linear is followed by SVR-RBF and XGBoost Tree, with a relative RMSEP of 145.8% and 147.7%, respectively. However, despite presenting a higher accuracy, XGBoost Linear tends to overfit the data, having an RMSEP to RMSEC ratio of 12.5, while SVR-RBF has a RMSEC to RMSEP ratio of 1.5.

The parity plots in figure 3.12 show the results of models after log transformation. The log transformation improves the accuracy of the models, but difficulties persist in accurately predicting the IP of iso-paraffins, particularly highly branched structures like 2,2,4-trimethylpentane, 2,2,4,6,6-pentamethylheptane, and 2,2,4,4,6,8,8-heptamethylnonane. We attribute this problem to the non-linear relationship between the number of quaternary carbons and the IP and the poor representation of these compounds in the database (see figure 2.12). Another major challenge for the accurate prediction of iso-paraffin stability is the limited number of compounds in the database compared to the vast number of possible isomers. For example, there are 618 045 structural isomers in the C_5-C_{20} carbon range [247]. Modeling limitations caused by the database size and class imbalance can be partially mitigated by performing DA

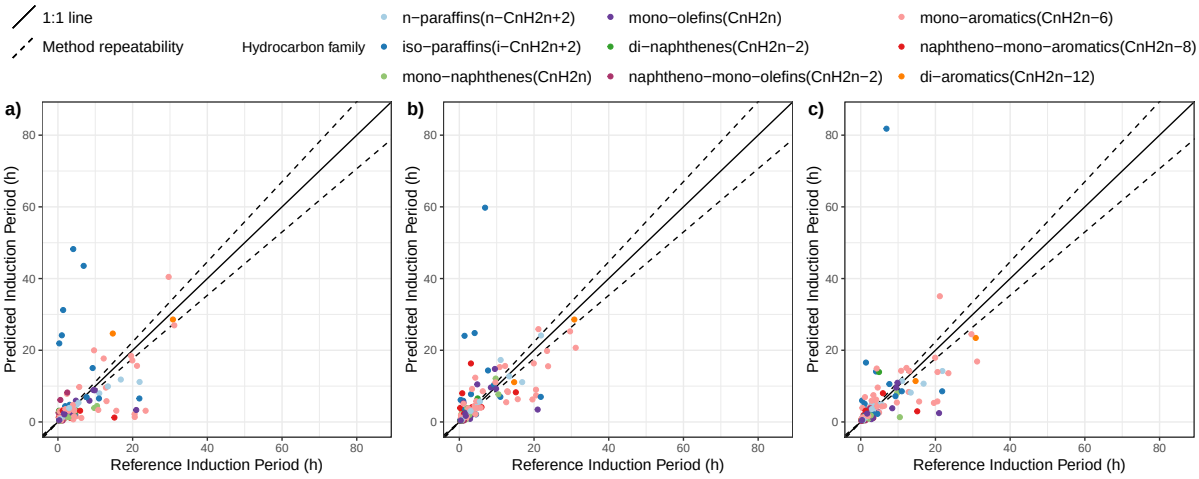


FIGURE 3.12 – Back-transformed reference vs. predicted IP values for the 4 external CV folds for models obtained using log(IP), without Data Augmentation. Shown for **a)** Support Vector Regression, **b)** XGBoost with linear learners, and **c)** XGBoost with regression trees.

As shown in table 3.5, data augmentation systematically improved the performance of all three algorithms by reducing the prediction error for underrepresented classes and iso-paraffins. For models trained without log-transformed data, RMSEP was reduced by 30% to 40% with respect to their non-augmented counterparts, with XGBoost Tree being the best-performing algorithm.

Figure 3.13 illustrates the results of the models trained for the prediction of the IP in h, after Data Augmentation. The use of Data Augmentation significantly improves the accuracy of the models for all the hydrocarbon families, including iso-paraffins. However, the models still present limitations for the prediction of low IP values due to the data skewness. Figure 3.14 presents the results of the models incorporating log transformation following Data Augmentation. Among them, the XGBoost Linear model, utilizing the combination of log transformation and Data Augmentation, demonstrates the best overall performance. Although DA improves the accuracy of the models, the prediction error of the best model remains relatively high (RMSEP = 2.67 h, i.e., an average relative error of 48.6%). This error is well above the repeatability of the reference method (11.7%). Consequently, the model can be considered a semi-quantitative tool, useful for initial hydrocarbons screening but not for precise quantitative prediction.

Table 3.5 – Model performance metrics for SVR-RBF, XGBoost with linear learners and XGBoost with regression trees learners applied on the data set resulting from DA. RMSE values were averaged across the 4 external cross-validation folds.

Model	**Optimal hyper-parameters**	**Scale**	**RMSEC**	**RMSECV**	**RMSEP**
SVR-RBF	$\sigma = 0.25$ $C = 100.00$ $\epsilon = 0.00$	hour	3.76	6.71	6.84
XGBoost Linear	n_rounds = 250 eta = 0.05 $\lambda = 2$ $\alpha = 15$	hour	1.63	6.65	4.67
XGBoost Tree	n_rounds = 150 eta = 0.2 max_depth = 8 $\gamma = 0.66$ colsample_bytree = 1 min_child_weight=0.25 subsample = 0.33	hour	2.32	6.09	4.36
SVR-RBF log-transform	$\sigma = 0.006$ $C = 10$ $\epsilon = 0.1$	log hour	0.48 4.24	1.24 5.87	0.88 5.10
XGBoost Linear log-transform	n_rounds = 750 eta = 0.05 $\lambda = 1$ $\alpha = 0$	log hour	0.17 1.59	1.21 3.84	0.67 2.67
XGBoost Tree log-transform	n_rounds = 200 eta = 0.3 max_depth = 6 $\gamma = 0$ colsample_bytree = 0.66 min_child_weight=0.5 subsample = 1	log hour	0.22 2.14	1.10 4.15	0.78 3.42

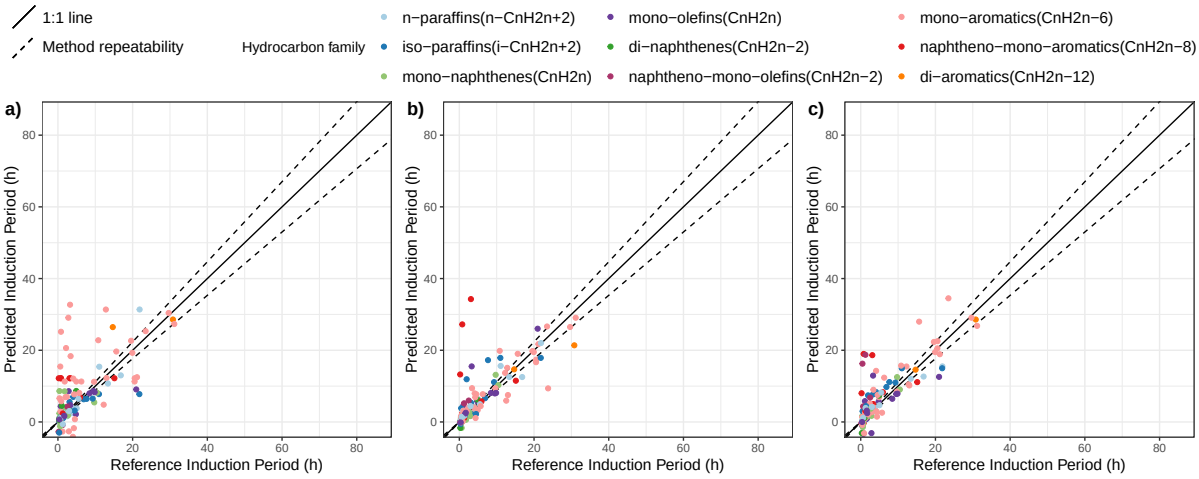


FIGURE 3.13 – Reference vs. predicted IP values for the 4 external CV folds after Data Augmentation. Shown for **a)** Support Vector Regression, **b)** XGBoost with linear learners, and **c)** XGBoost with regression trees.

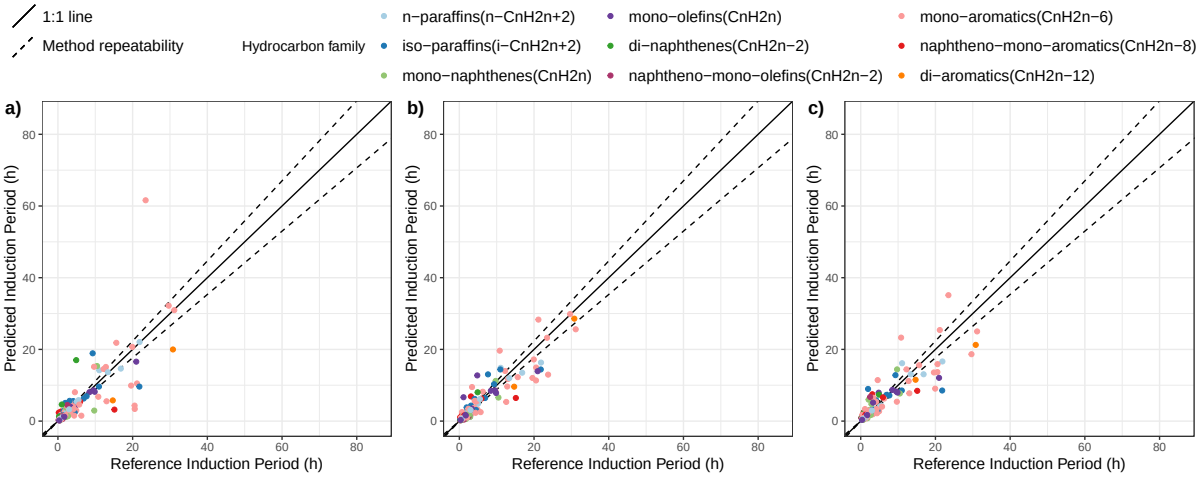


FIGURE 3.14 – Back-transformed reference vs. predicted IP values for the 4 external CV folds for models obtained using log(IP), after Data Augmentation. Shown for **a)** Support Vector Regression, **b)** XGBoost with linear learners, and **c)** XGBoost with regression trees.

3.4.2 Prediction of Induction Period trends

The capacity to predict the IP of hydrocarbons through ML-QSPR methods also offers a valuable tool for both validation and the analysis of trends. In this section, predictions generated using the best model (XGBoost Linear) together with the available experimental data are used to analyze four evolution trends for IP with some molecular features: the effect of the paraffin carbon number, the influence of linear side-chain length on mono-aromatic compounds, the impact of the methyl group substitutions on an aromatic ring, and the number of carbon atoms in 1-olefins.

Figure 3.15**a** shows that the Induction Period of paraffins is negatively affected by the carbon atom number. Furthermore, the model is able to capture said trend, where the IP rapidly decreases until carbon number 10, and then reaching a plateau at about 2.5 hours [90, 248].

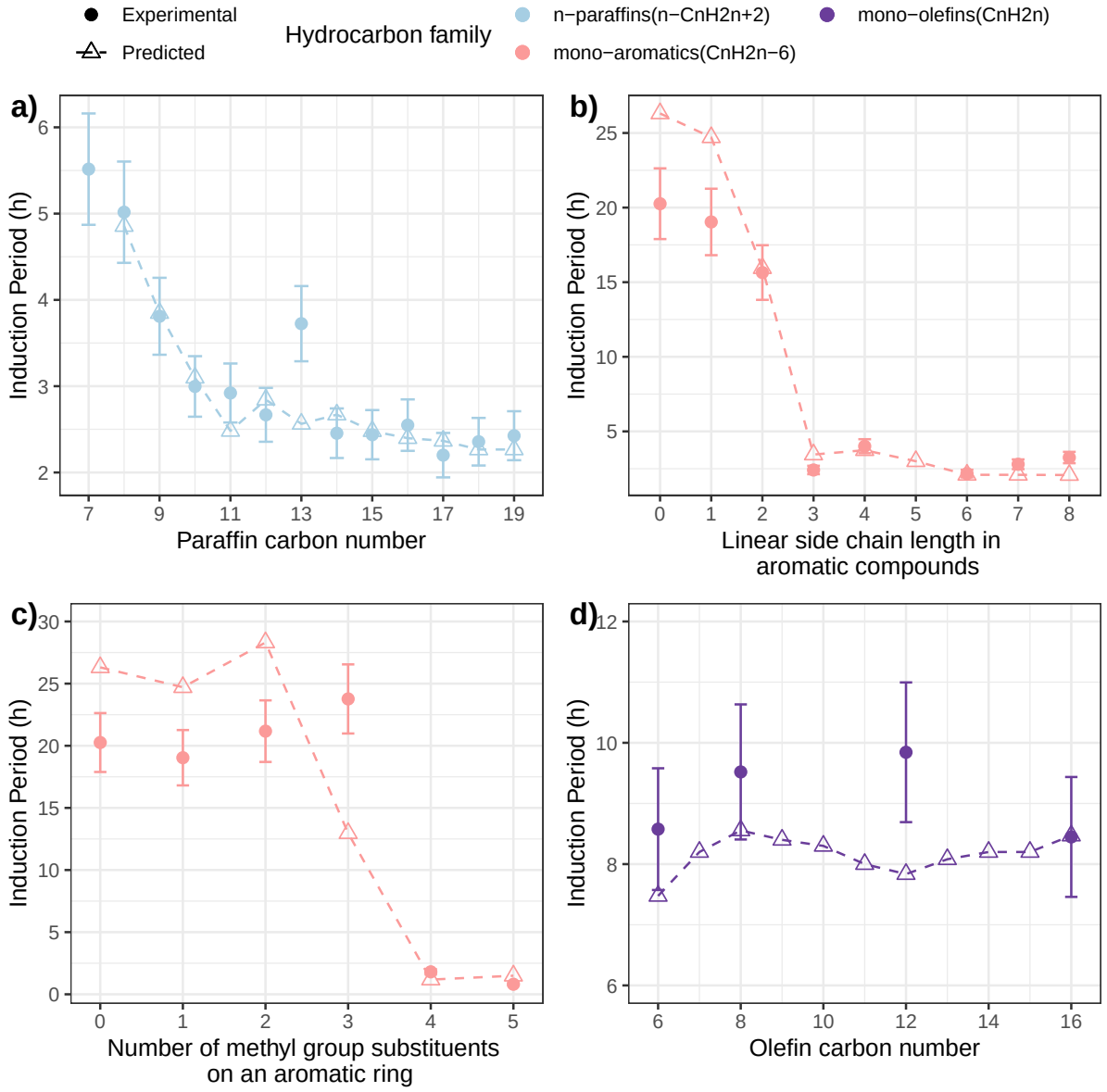


Figure 3.15 – Experimental and predicted trends using the XGBLinear model, trained on log-transformed data and with DA. **a)** Induction Period as a function of the carbon number in linear paraffins. **b)** Induction Period as a function of the linear side chain length in mono-aromatic compounds. **c)** Induction Period as a function of the number of methyl substituents on an aromatic ring, for benzene, toluene, *m*-xylene, mesitylene, durene and pentamethylbenzene. **d)** Induction Period as a function of the carbon number in 1-olefins. IP values are reported at 140 °C, except for olefins, which are reported 100 °C.

Interestingly, the model doesn't capture the deviation from the trend for *n*-tridecane, suggesting that the model is capable to overcome this sort of experimental artifacts. Figure 3.15**b** illustrates the negative correlation between the Induction Period and the linear side chain

length in aromatic compounds. For benzene and toluene, the model overestimates the IP by approximately 25%. However the accuracy of the model predictions increases for aromatic compounds with longer side chains. Noting that for side chains composed of 3 carbon atoms and more, IP values reach a plateau at about 2.5 hours. Figure 3.15**c** presents the effect of the number of methyl substituents on the IP of mono-aromatic compounds. For molecules containing less than 3 methyl group substituents, the IP is not negatively affected, with IPs roughly equal to 20 hours. However, the stability decreases when 4 or more substituents are present, with IP of about 2 hours. Our model is able to reproduce this general trend, presenting a higher prediction error for compounds with less than 4 substituents. Figure 3.15**d** shows the relationship between IP values and carbon atoms number in 1-olefins. As noted in Venegas-Reynoso et al. [248], the variation in IP with carbon number is not statistically significant when considering the method repeatability. The model effectively reproduces the experimental trend across the C_6 to C_{16} range and provides estimates for missing values in the literature.

3.4.3 Model interpretation

In addition to making predictions, another key objective of modeling is to gain insights into the relationships between input features and the studied phenomena. Although the accuracy of our model remains limited at some points, its interpretation enables us to examine the connections between molecular features and oxidation stability and contrast these results with real data. The SHAP values and mean absolute SHAP values for the top-performing model, XGBoost Linear with log-transformation and DA, are shown in figure 3.16**a** and figure 3.16**b**, respectively. For clarity, model features are encoded as presented in table 3.3. Next, according to their mean absolute SHAP value, we will talk about the ten most descriptive traits. Temperature (T), which shows a negative correlation with the Induction Period, is the most significant feature. The literature has provided substantial documentation of this trend [50, 90, 125].

The results show that an approximate doubling of the Induction Period occurs when the temperature is reduced by 10 K [62]. The number of allylic carbons with at least one hydrogen in non-conjugated double bonds (X18), which also adversely affects the IP, is the next crucial characteristic. Because of the strong reactivity linked to the generation of resonance-stabilized allylic radicals, olefins are recognized for having low oxidation stability [89, 159]. This outcome is consistent with our earlier research [248], which found that, as long as the substituting carbon atom keeps at least one hydrogen atom, olefin reactivity positively correlates with the extent of hydrogen substitution at vinylic sites. The amount of secondary carbons bonded to aliphatic carbons that are not in a ring is the third most important molecular descriptor (X4). Since hydrocarbon stability, such as that of linear alkanes or alkyl-aromatics, decreases with the length of the paraffinic chain, this molecular property is inversely connected with

the Induction Period [89, 90, 248]. The molecular mass (X14) is the fourth most significant attribute. It has a strong correlation with the sixth-most significant feature, the total amount of hydrogen atoms (X1). The oxidation stability has a positive correlation with both features. The IP values of *n*-hexane (7.5 h) and *n*-eicosane (2.1 h) in the linear alkanes family suggest that larger molecules are generally less stable [90]. There are, however, certain exceptions; compounds with large molecular masses, like biphenyl and methyl-naphthalene isomers, have high IP values.

The number of primary carbon atoms in non-cyclic paraffins (X2), which shows a positive link with the IP, is the fifth most significant property. The great stability of quaternary carbons, which are frequently found in heavily substituted compounds in our sample, may be the cause of this trend. In contrast, the preponderance of tertiary carbon atoms in other highly substituted compounds, like 2,3,4-trimethylpentane, results in reduced stability. The number of tertiary C-atoms linked to an aromatic ring (X11) that are not in a ring, like the benzylic centers in cumene, p-cymene, 1,4-diisopropylbenzene and 1,3,5-triisopropylbenzene, is the seventh most important feature. This quantity exhibits a negative correlation with the IP. Compared to the C-H bond-dissociation energies of primary and secondary C atoms attached to aromatic rings, tertiary benzylic centers exhibit heightened reactivity due to increased substitution. For instance, IP of cumene is approximately 80 and 65 times than those of toluene and ethylbenzene, respectively. When secondary carbon atoms within a ring are joined to an unfused aromatic ring (X35), as is the situation with the benzylic carbon atoms in tetralin, the result is comparable but opposite. Because of the ring strain instability and higher reactivity associated with benzylic sites, this characteristic has a negative correlation with the IP [248]. The next critical feature is the number of aromatic carbon atoms (X15), which present a positive correlation to the IP. Because benzene's C-H bond-dissociation energy produces an unstable phenyl radical, aromatics are highly stable [249]. Naphthalene and biphenyl likewise show this trend, with high IP values of 33.5 and 25.4 hours, respectively [248]. The number of quaternary carbon atoms connected to unfused aromatic rings (X13), which includes stable compounds like tert-butylbenzene and tert-pentylbenzene, is the tenth most significant feature. The lack of H-atoms in the benzylic site, which are frequently engaged in hydrogen abstraction, is what gives these compounds their stability [50, 248].

On the other hand, some features have minimal impact on the model predictions. This is particularly true for features present in only one sample, such as the number of tertiary carbons in a ring adjacent to an aromatic carbon (X32), for example, in 1,5-dimethyltetralin, or the number of secondary carbons attached to a naphthalene ring (X42). Another case of features considered unimportant is when two similar molecules, for instance, 1- or 2-methylnaphthalene, and naphthalene, have similar IP values. Thus, the number of primary carbons attached to the naphthalene ring (X40) has a SHAP value close to 0. Therefore, the interpretation of our

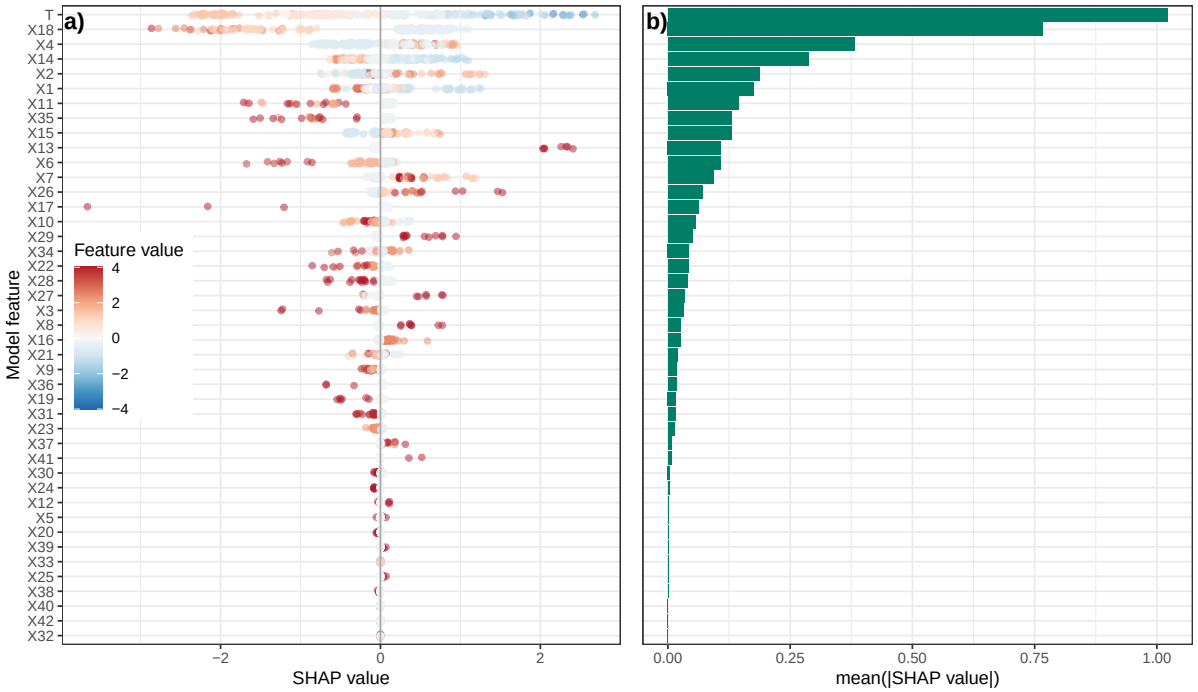


FIGURE 3.16 – SHAP values and mean absolute SHAP values for the features of the best-performing model: XGBoost Linear, with log-transformation and DA.

model's SHAP values depends on the dataset and may change as additional molecules are included.

3.5 Conclusions

ML-QSPR based models provide a fast and accurate alternative to the traditional kinetic modeling approaches for the determination of molecular physico-chemical properties. In this work, we presented the first attempt to develop ML-QSPR models for the prediction of the Induction Period for pure liquid phase hydrocarbons using Support Vector Regression with the Radial Basis Function kernel (SVR-RBF), and eXtreme Gradient Boosting with tree (XGBoost Tree) and linear (XGBoost Linear) learners. We also studied the effect of log-transforming the reference Induction Period and applying Data Augmentation, on the predictive power of the models. We observed that despite presenting overfitting, XGBoost outperformed SVR-RBF, while XGBoost Linear was the best-performing model. Furthermore, applying data augmentation to the dataset reduced the model's prediction error by 36 to 60% depending on the algorithm. Thus, the overall best-performing model we obtained is XGBoost Linear for log-transformed IP values and data augmentation, with RMSEP = 2.67 h, which represents a relative error of 48.6%. While the model's accuracy is not adequate for quantitative predictions when compared to the reference method's error (approximately 10%), it permits its use as a semi-quantitative model. On the other hand, the calculation of SHAP values for model

interpretation allows us to understand at a global level, the most relevant molecular features for liquid phase oxidation reactivity and their correlation to the Induction Period. At a local level, SHAP values help to understand which features are responsible for the predicted values of a given sample. ML-QSPR models, combined with the analysis of key molecular features influencing hydrocarbon oxidation stability, can serve as an effective screening tool in the development of Sustainable Aviation Fuel. Future work may focus on the improvement of model accuracy, by increasing the dataset size, investigating new molecular descriptors, and possibly, supplementing said descriptors with other chemical information, such as spectral data.

Chapter 4

NIR spectroscopy-based modeling of the oxidation stability of hydrocarbons

4.1 Introduction

In the previous chapter, we explored a data-driven modeling approach known as Quantitative Structure-Property Relationship (QSPR), a key area within cheminformatics. We examined the process of encoding molecular structures into vector representations, developed a set of molecular descriptors specifically relevant to oxidation stability, and presented the results of the generated models. The chapter concluded with a model interpretation step, highlighting the most significant descriptors influencing oxidation stability. Additionally, we discussed how data-driven modeling typically functions as a "black-box," converting inputs into outputs without need of internal mechanistic knowledge. While QSPR modeling utilizes molecular representations as inputs, this chapter shifts focus to a different type of input: spectral data.

The use of spectral data as input for various types of models is a characteristic of chemometrics, another field that integrates chemistry with statistics and multivariate analysis. In this chapter, we present a new data-driven study based on the accelerated oxidation tests discussed in Chapter 2 and spectral information of the samples in our database. For this, we will first discuss different spectroscopic techniques, then we will provide a brief introduction to the field of chemometrics and proceed to the modeling step and discussion of the modeling results.

4.2 Spectroscopic techniques

Spectroscopy is an analytical technique used to study the interaction of matter with electromagnetic radiation. The analysis of the different types of radiation-matter interactions; absorption, emission, and scattering across various wavelengths provides valuable insights

into the structure, composition, and properties of materials [250]. Said interactions can be understood as follows:

- **Absorption:** When an atom, ion, or molecule transitions from a lower energy state to a higher energy state, it absorbs photons whose energies match the energy gap between these states [250]. For example, for a generic molecule, the absorption can be written as:

$$\mathrm{M} + h\nu \longrightarrow \mathrm{M}^* \tag{4.1}$$

  where M is a generic molecule in ground-state, $h\nu$ the energy of photon expressed as the product of Planck's constant and ν the frequency of the radiation, and M^*, the molecule in its excited state.

- **Emission:** The process where a chemical species in a higher energy state loses energy, emitting a photon.

$$\mathrm{M}^* \longrightarrow \mathrm{M} + h\nu \tag{4.2}$$

- **Scattering:** Scattering is the process by which electromagnetic radiation is redirected or deflected as it interacts with particles on molecules. There are two main types of scattering: elastic or Rayleigh scattering, and inelastic or Raman scattering. In Rayleigh scattering, the both deflected and incident photons have the same wavelength ($\nu_1 = \nu_2$):

$$\mathrm{M} + h\nu_1 \longrightarrow \mathrm{M} + h\nu_2. \tag{4.3}$$

  In contrast, in Raman scattering, the deflected and incident photons have different wavelengths due to energy transfer to the molecule. It is estimated that about 1 in 10^7 of the deflected photons undergo Raman scattering, making it a relatively rare phenomenon [251, Chapter 13]. Moreover, this type of scattering can be further divided into two categories [252]. The first is Stokes Raman scattering; when a molecule gains energy and moves from the ground vibrational state to an excited vibrational state, thus the scattered photon has lower energy ($\nu_1 > \nu_2$):

$$\mathrm{M} + h\nu_1 \longrightarrow \mathrm{M}^* + h\nu_2. \tag{4.4}$$

  The second type of inelastic scattering is anti-Stokes Raman scattering; when a molecule loses energy and moves from an excited vibrational state to the ground state, thus, the deflected photon has higher energy ($\nu_1 < \nu_2$):

$$\mathrm{M}^* + h\nu_1 \longrightarrow \mathrm{M} + h\nu_2, \tag{4.5}$$

The core of spectroscopy is the electromagnetic spectrum, a continuous range of wavelengths and frequencies of electromagnetic radiation. The spectrum can be divided into regions based on the wavelength and energy of the radiation. These regions include gamma-rays, X-rays, ultraviolet (UV), visible light, infrared (IR), microwaves, and radio waves, as shown in figure 4.1 [253].

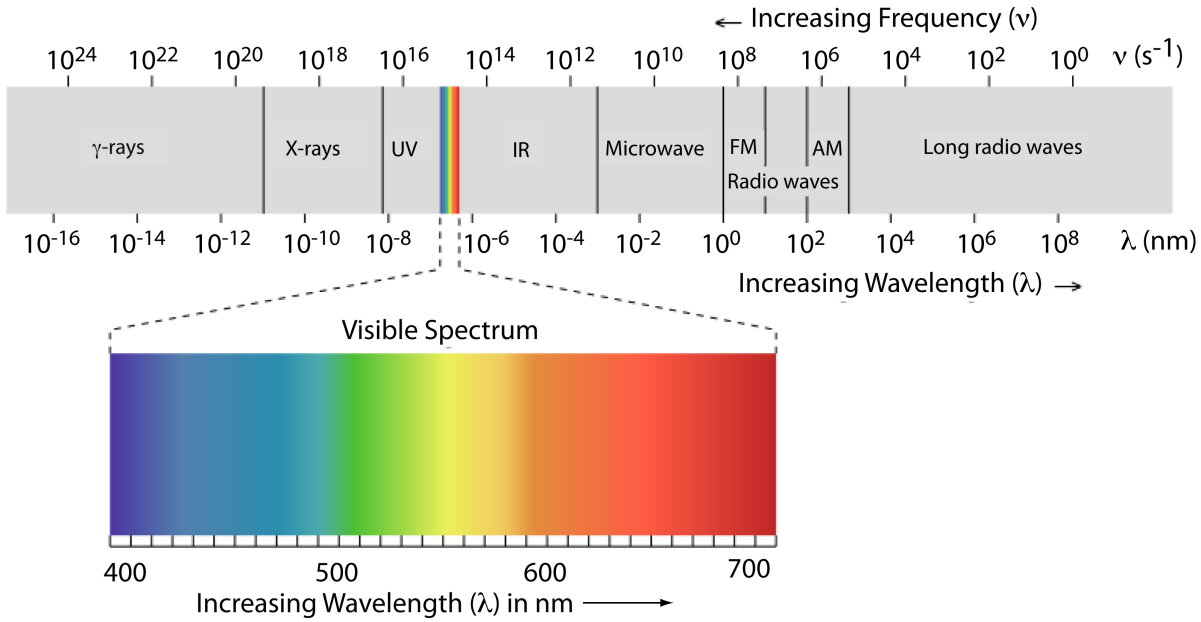


FIGURE 4.1 – The electromagnetic spectrum and the boundaries between its different regions. Reproduced from LibreTexts [253].

Each region of the electromagnetic spectrum produces different and specific radiation-matter interactions. For instance, gamma-rays cause nuclear transitions, while X-rays are linked to inner shell electron transitions, useful for elemental analysis. Ultraviolet and visible light primarily induce valence electron transitions, which help to understand electronic structures. In contrast, infrared radiation is closely related to molecular vibrations, and microwaves correspond to molecular rotations. Radio waves, on the other hand, are employed in nuclear magnetic resonance (NMR) spectroscopy, which study nuclear spin transitions [253]. Thus, various methods have been developed to exploit these interactions for the analysis of matter. A summary of the electromagnetic spectrum regions, the types of interactions they produce, and the corresponding techniques is provided in table 4.1.

In this work, we will focus on molecular spectroscopic techniques that are useful for the analysis of hydrocarbons. Thus, we will omit Mössbauer, X-ray, and atomic absorption spectroscopies, since they are mainly used for elemental analysis.

TABLE 4.1 – Absorption and scattering spectroscopic techniques across the electromagnetic spectrum. Adapted from LibreTexts [253].

Electromagnetic spectrum region	Type of atomic or molecular transition	Spectroscopic technique
Gamma-rays	Nuclear	Mössbauer spectroscopy
X-rays	Core-level electrons	X-ray absorption spectroscopy
UV/Vis	Valence electrons	UV/Vis spectroscopy
		Atomic absorption spectroscopy
IR	Molecular vibrations	Infrared spectroscopy
		Raman spectroscopy
Microwave	Molecular rotations	Microwave spectroscopy
	Electron spin	Electron spin resonance spectroscopy
Radio wave	Nuclear spin	Nuclear magnetic resonance spectroscopy

4.2.1 UV-Vis spectroscopy

UV-Vis spectroscopy operates on the principle of light-matter interactions within the ultraviolet ($\approx 180 - 400$ nm) and visible ($\approx 400 - 700$ nm) spectral regions. At these wavelengths, valence electron transitions occur, involving the promotion of electrons from the Highest Occupied Molecular Orbital (HOMO) to the Lowest Unoccupied Molecular Orbital (LUMO). As shown in figure 4.2, in sigma bonds, the transition is denoted as $\sigma \rightarrow \sigma^*$. Likewise, electrons from pi-bonding orbitals can be promoted to anti-bonding pi orbitals ($\sigma \rightarrow \sigma^*$). Similarly, lone electron pairs in non-bonding orbitals (n) from heteroatoms can be excited to anti-bonding pi (n $\rightarrow \pi^*$) orbitals, or anti-bonding sigma (n $\rightarrow \sigma^*$) orbitals [254].

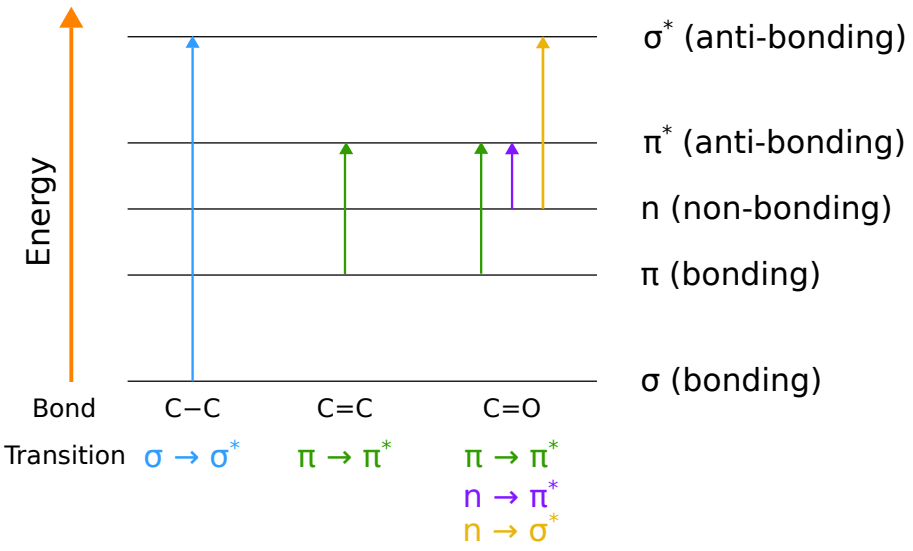


FIGURE 4.2 – Electronic transitions in the UV-Vis region. Adapted from LibreTexts [255].

UV-Vis spectroscopy is a type of absorption spectroscopy, where the spectrometer measures the ratio between the intensity of light emitted by the source and the intensity of light after it passes through the sample. This ratio is directly related to the sample concentration according to the Lambert-Beer law [256]:

$$A = \log_{10}\left(\frac{I_0}{I}\right) = \epsilon l c \tag{4.6}$$

where:

- A is the absorbance, which is defined by the incident intensity I_0 and transmitted intensity I,
- ϵ, the molar extinction coefficient, which depends on the sample,
- l, the length of the light path, and
- c, the analyte concentration.

In the context of fuel stability, UV-Vis has been used for the quantification of antioxidants in biodiesel [257]. Although this spectroscopic technique allows for accurate compound quantification, it does not provide specific structural information, such as the identification of bonding patterns, rendering this technique unsuitable for the identification of useful molecular patterns related to stability.

4.2.2 Infrared spectroscopy

IR spectroscopy is type of absorption spectroscopy, widely used analytical technique based on the interaction of infrared radiation with matter. Molecules absorb radiation in the IR region of the electromagnetic spectrum, causing vibrational transitions that lead to the deformation of molecular bonds. These bond deformations can occur as stretching (changes in bond length) or bending (changes in bond angles), depending on the energy and mode of vibration. For a molecule to be IR-active, the vibration must result in a change in the dipole moment of the molecule [258].

The IR region of the electromagnetic spectrum is typically divided into three subregions [259, Chapter 1]:

- **Near-Infrared (NIR):** 12,821 to 4,000 cm^{-1} (780 to 2,500 nm). This region is associated with overtones and combination bands of fundamental vibrations.
- **Mid-Infrared (MIR):** 4,000 to 400 cm^{-1} (2,500 to 25,000 nm). The most commonly used region in IR spectroscopy, where fundamental vibrational transitions occur.
- **Far-Infrared (FIR):** 400 to 10 cm^{-1} (25,000 to 1,000,000 nm). This region is associated with low-energy vibrations, such as torsional and lattice modes in solids.

The behavior of molecular vibrations can be approximated using the harmonic oscillator model (see figure 4.3). In this model, two atoms in a bond are treated as masses connected by a spring, where the bond acts as a restoring force. The harmonic oscillator assumption leads to quantized vibrational energy levels, described by:

$$E_{\text{VIB}} = h\nu \left(\upsilon + \frac{1}{2} \right) \tag{4.7}$$

where:

- E_{VIB} is the discrete vibrational energy levels,
- h, Planck's constant,
- ν, the vibrational frequency of the bond, and
- υ, the vibrational quantum number that takes integer values 0, 1, 2, *etc.*

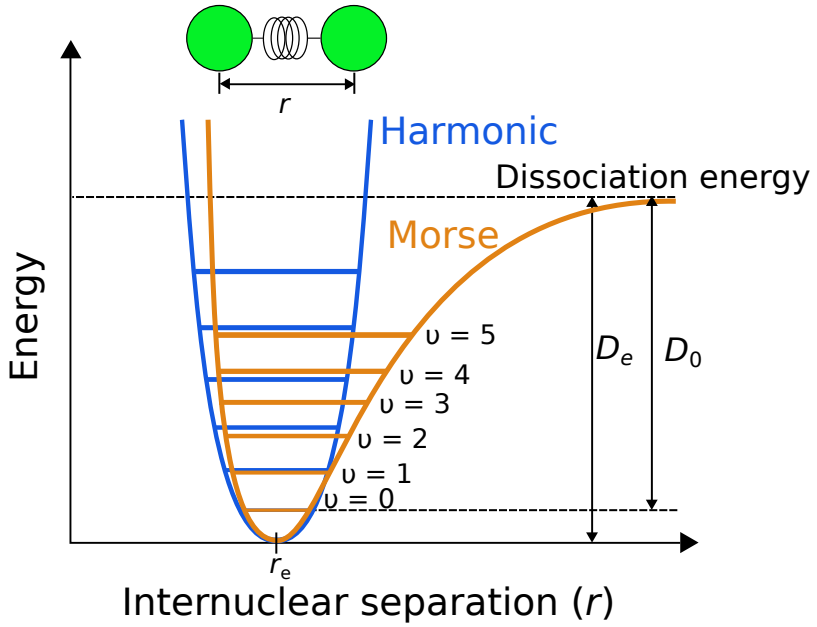


FIGURE 4.3 – Comparison of the harmonic oscillator and Morse potential energy curves as a function of internuclear separation (r). The harmonic potential (blue curve) assumes equally spaced energy levels, while the Morse potential (orange curve) better represents molecular vibrations, accounting for anharmonicity and dissociation energy. D_e represents the depth of the potential well, D_0 is the dissociation energy accounting for zero-point energy, and r_e is the equilibrium bond length. Vibrational quantum numbers (υ) are shown for the Morse potential. Adapted from LibreTexts [260].

While the harmonic oscillator model provides a good approximation, real molecular vibrations deviate from this behavior due to anharmonicity. For instance, the harmonic oscillator does not consider bond-dissociation as the inter-nuclei distance increases. In the anharmonic oscillator model, the energy levels are not equally spaced, as the potential energy well flattens near bond dissociation. Furthermore, overtones (transitions beyond the fundamental, such as $\upsilon = 0 \rightarrow 2$) and combination bands become possible, leading to additional peaks in the IR spectrum. Also, an anharmonicity constant, x_e, for each bond, modifies the vibrational

energy levels, improving the representation of real molecular behavior. Anharmonicity can be described using interatomic interaction models for the potential energy, such as the Morse potential [259, Chapter 1].

In this work, we will omit the discussion of the Far-Infrared (FIR) region, as it is primarily used for the analysis of materials, peptides, proteins, and gases [261].

4.2.2.1 Near-infrared spectroscopy

Near-Infrared spectroscopy is an analytical technique that utilizes the near-infrared region of the electromagnetic spectrum, covering wavelengths from approximately 780 to 2500 nm [259, Chapter 1]. Its widespread applications in fields such as agriculture, pharmaceuticals, food science, and environmental monitoring are largely due to its non-destructive nature, and rapid analysis, often requiring minimal sample preparation [262].

When NIR radiation interacts with a sample, certain wavelengths are absorbed due to molecular vibrations. These vibrations are characteristic of chemical bonds, particularly those involving hydrogen atoms such as O-H, C-H, and N-H bonds. This makes NIR spectroscopy particularly effective for analyzing substances rich in water, proteins, and hydrocarbons [259, Chapter 1].

In the NIR region, absorption arises from overtone and combination bands, which are less intense than fundamental vibrational transitions ($\upsilon = 0$ to $\upsilon = 1$). Overtone bands arise from transitions that are integer multiples of fundamental vibrational frequencies. For example, when the vibrational level transitions from $\upsilon = 0$ to $\upsilon = 2$, it is referred to as the first overtone. Similarly, a transition from $\upsilon = 0$ to $\upsilon = 3$ corresponds to the second overtone, and so on. On the other hand, combination bands result from simultaneous excitation of two or more fundamental modes. These transitions produce broader and weaker absorption features compared to the fundamental bands, but they are rich in information regarding the chemical and physical properties of the sample [260].

The overlapping nature of NIR absorption bands can complicate spectral interpretation. However, this spectral technique is routinely used for quantitative analysis when combined with calibration models and chemometric techniques. Several studies involving NIR spectroscopy in the context of fuels have focused on diesel and biodiesel blends.

Recent studies have demonstrated the versatility of NIR spectroscopy in fuel analysis. For example, Velvarská et al. [263] employed NIR spectroscopy to predict the PetroOxy induction period in diesel/biodiesel blends. Similarly, Wang et al. [264] combined NIR with SVR to determine key fuel properties, including density, viscosity, and freezing point in diesel. More recently, Varghese et al. [265] utilized NIR spectroscopy with multilinear regression to assess various biodiesel oxidation properties, such as the conjugated diene, peroxide and iodine

values, degree of unsaturation, oxidative stability index, among others. In these works, the authors obtained satisfactory models for the determination of the aforementioned parameters. Furthermore, the authors highlight that the portability of NIR spectrometers and the coupling to optic fiber probes, make this technique a promising candidate for online analysis of fuels.

4.2.2.2 Mid-infrared spectroscopy

Mid-Infrared (MIR) spectroscopy is an analytical technique that studies the fundamental vibrational modes of molecules, typically within the wavelength range of (4000–400 cm^{-1}). Fundamental vibrations, such as stretching, bending, and torsional motions, occur when molecular bonds absorb specific frequencies of infrared radiation corresponding to their natural vibrational frequencies. These vibrations are unique to the chemical bonds and functional groups within a molecule [259, Chapter 1].

MIR spectroscopy provides detailed chemical information, including the identification of functional groups, molecular structures, and covalent bond types. For example, the technique can distinguish between single, double, and triple bonds, detect the presence of specific functional groups, such as hydroxyl, carbonyl or amine groups. Furthermore, this technique shows non-covalent interactions, such as hydrogen bonding. Additionally, MIR spectra are highly specific, serving as molecular fingerprints for qualitative and quantitative analysis of complex mixtures [258].

MIR spectroscopy has been used for the determination of carboxylic acid and phenol in biocrude [266], and the total acid number in petroleum [267]. Although not directly related to fuels, Wen et al. [268] developed a model to predict the oxidation stability of walnut oil. The study involved measuring the MIR spectra of fresh samples, followed by the analysis of the induction period using the Oxitest AOCS International Standard Procedure (Cd 12c-16). Through model interpretation, the authors identified key spectral bands associated with stability, including 3008, 1654, 914, and 723 cm^{-1}. These bands correspond to $=C-H$ and $C=C$ stretching, $=C-H$ out-of-plane bending, and $-CH_2-$ rocking vibrations, respectively [268].

4.2.3 Raman spectroscopy

Raman spectroscopy, a branch of vibrational spectroscopy, enables the structural identification and quantification of molecules by analyzing their unique vibrational fingerprints [269]. The Raman effect, predicted by Smekal in 1923, was first observed experimentally by Raman and Krishnan in 1928. During their experiments, one of Raman's students discovered that when sunlight filtered through violet glass passed through purified water and alcohol, the scattered rays exhibited wavelengths different from those of the incident beam [252].

As previously mentioned, this wavelength shift can occur in either direction, resulting in Stokes and anti-Stokes Raman scattering (see figure 4.4). The intensity of anti-Stokes Raman lines is generally weaker than that of Stokes lines, as their ratio is governed by the Boltzmann distribution. Consequently, anti-Stokes scattering is less commonly used for analytical purposes [252].

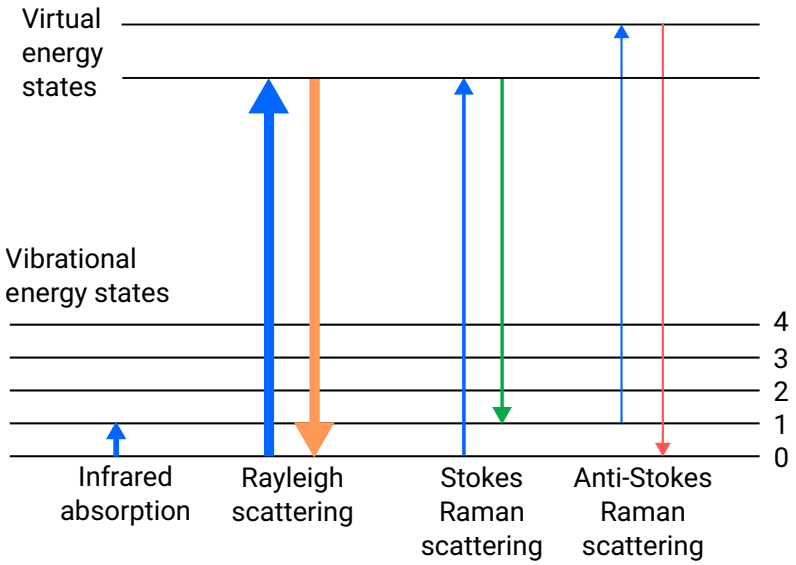


FIGURE 4.4 – Energy diagram illustrating infrared absorption, Rayleigh scattering, and Stokes and anti-Stokes Raman scattering. Infrared absorption involves direct transitions between vibrational states. Rayleigh scattering is elastic, with no change in vibrational energy. Stokes scattering transitions to higher vibrational states, while anti-Stokes transitions to lower states, resulting in longer or shorter scattered wavelengths, respectively. Adapted from LibreTexts [270].

The intensity of a Raman signal is directly proportional to the magnitude of the change in molecular polarization, while the wavelength shift of the scattered light provides information about the molecular structure of the compounds [269].

Raman spectroscopy has been applied in several studies related to fuel analysis. These include the discrimination and quantification of biodiesel in blends with fossil diesel [271], the determination of Reid vapor pressure as well as motor and research octane numbers in petroleum fuels [272], and the estimation of cetane numbers in jet fuel mixtures [273].

4.2.4 Nuclear Magnetic Resonance (NMR) and Electron Spin Resonance (ESR) spectroscopy

Nuclear Magnetic Resonance (NMR) and Electron Spin Resonance (ESR) are powerful spectroscopic techniques used to study molecular and electronic properties. Both rely on the interaction of magnetic fields with magnetic moments associated with nuclei or unpaired

electrons. Below, we discuss the key principles, transitions involved, and the information provided by these methods.

NMR spectroscopy exploits the magnetic properties of certain nuclei, such as ^{1}H and ^{13}C, which possess a nonzero spin. In the presence of an external magnetic field (B_0), these spins align in either a lower-energy state (parallel to B_0) or a higher-energy state (antiparallel to B_0), a process known as Zeeman splitting (see figure 4.5) [274]. The energy difference (ΔE) between these states is given by:

$$\Delta E = \hbar \gamma B_0, \tag{4.8}$$

where:

- γ is the gyromagnetic ratio of the nucleus, and
- $\hbar$, the reduced Plank constant, equal to $h/2\pi$.

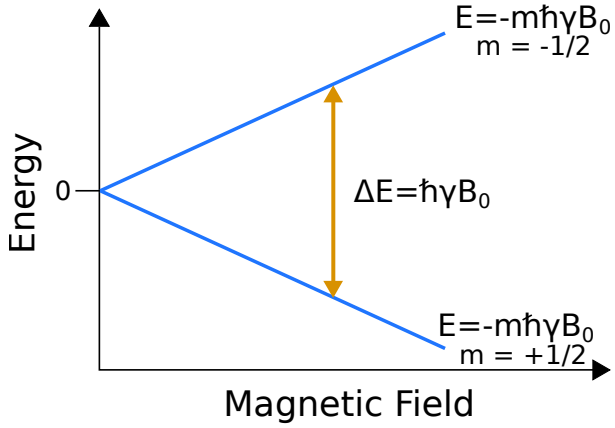


FIGURE 4.5 – Zeeman splitting of a spin-$\frac{1}{2}$ nucleus in a magnetic field (B_0), with an energy gap $\Delta E = \hbar \gamma B_0$ corresponding to NMR transitions. Adapted from LibreTexts [274].

Absorption of radiofrequency radiation matching ΔE causes transitions between the spin states, generating the NMR signal. However, the chemical environment around a nucleus influences its local magnetic field through electron shielding. This results in shifts in the resonance frequency relative to a standard reference, such as tetramethylsilane (TMS). These "chemical shifts" provide information about functional groups, bond types, and molecular structure. Chemical shifts are reported in parts per million (ppm), denoted as δ, which are calculated as follows [275]:

$$\delta = \left(\frac{H_{\text{ref}} - H_{\text{sub}}}{H_{\text{machine}}} \right) \times 10^6 \tag{4.9}$$

where:

- H_{ref} is the resonance frequency of the reference,

- H_{sub}, the resonance frequency of the substance, and,
- H_{machine}, the resonance frequency of the spectrometer.

The use of chemical shift allows for direct comparison between experiments and systems. While the chemical shift helps identify the types of functional groups in a molecule, the finer details of the NMR spectrum arise from spin-spin coupling interactions. Spin coupling occurs due to the magnetic influence of neighboring nuclei, splitting resonance peaks into multiplets. For spin-$\frac{1}{2}$ nuclei, the splitting pattern follows Pascal's triangle:

- A single neighboring nucleus produces a doublet with a 1:1 intensity ratio.
- Two equivalent neighboring nuclei create a triplet with a 1:2:1 ratio.
- Three equivalent nuclei yield a quartet with a 1:3:3:1 ratio.

This splitting pattern, combined with the chemical shift, reveals relevant structural information, such as the number of adjacent nuclei and their connectivity [275].

Given its ability to identify molecular features, such as bonding patterns and aromaticity, NMR is a valuable tool for modeling material properties. For example, Kumar et al. [276] utilized ^{1}H-NMR spectroscopy to determine the iodine value in biodiesel. In their study, the authors analyzed the signal from the OCH_3 and α-CH_2 groups at 3.67 and 2.2 ppm, respectively, to quantify the methyl ester content. Thus, the authors indirectly assessed the oxidation stability of the samples, as it is inversely correlated to the iodine value.

Electron Spin Resonance (ESR) spectroscopy, also called Electron Paramagnetic Resonance (EPR) spectroscopy, targets unpaired electrons, which have a spin quantum number $S = \frac{1}{2}$ [277]. When subjected to an external magnetic field, the electron spin undergoes Zeeman splitting, which results in two energy states, $m_s = +\frac{1}{2}$, and $m_s = -\frac{1}{2}$. The energy difference for these two states, ΔE, is given by the equation [278]:

$$\Delta E = g\mu_B B_0, \tag{4.10}$$

where:

- g is the g-factor, which is specific to the electronic environment, and roughly equivalent to the chemical shift in NMR spectroscopy,
- μ_B, the Bohr magneton constant, and
- , the external magnetic field.

Transitions between these states occur upon absorption of microwave radiation matching ΔE when there are unpaired electrons in the system. Interactions between unpaired electrons and nearby nuclei cause "hyperfine splitting" in the ESR spectrum. This splitting provides detailed information about the electronic structure and local environment, such as the number of nuclei interacting with the unpaired electron and their spin states [279].

ESR spectroscopy is a powerful tool for studying autoxidation processes due to its ability to directly detect the formation of radicals. For example, Andersen [280] employed this technique to investigate lipid oxidation, while Babić et al. [281] examined the formation of free radical intermediates in polyalphaolefin base oil. However, because the radicals in these systems are highly reactive and short-lived, both studies utilized spin-trapping methods to stabilize and detect the radicals effectively.

4.3 Chemometrics

The term "chemometrics" was introduced by Svante Wold in 1974, who defined it as "the art of extracting chemically relevant information from data produced in chemical experiments" [282]. This multidisciplinary field combines statistical, mathematical, and computational approaches to optimize experimental designs and measurement procedures. It aims to maximize the extraction of chemical information from experimental data [283]. Some examples of problems that concern chemometrics include multivariate calibration, sampling theory, time series analysis, pattern recognition, data reduction, experimental design, signal deconvolution, image analysis, multivariate curve resolution, among others [284].

In this thesis, we discuss three topics closely related to our objective of developing a model based on NIR spectra; pre-processing, dimensionality reduction, and regression techniques.

4.3.1 Pre-processing

Spectroscopic data often contain undesirable artifacts such as noise, baseline drifts, and physical effects that obscure chemically relevant information. Pre-processing is an essential step to minimize the influence of these artifacts, isolate the chemical signal, and improve the quality of subsequent models. For example, in Near-Infrared Spectroscopy (NIRS), light scattering alters the path-length of the light beam traveling from the sample to the detector, which in turn changes signal intensity and lowers the predictive accuracy of models based on raw data.

Common artifacts encountered in spectroscopic data include additive and multiplicative effects, response curvatures, wavelength shifts, and random heteroscedastic noise. These effects can be expressed as:

$$\mathbf{z} = f(\mathbf{z}_{\text{true}}) = b\mathbf{z}_{\text{true}} + a, \tag{4.11}$$

where:

- $\mathbf{z}$ is the experimentally measured spectrum,
- $\mathbf{z}_{\text{true}}$, the spectrum containing only chemical information,

- a, a constant offset, and
- b, a scaling factor.

To address these challenges and ensure that spectroscopic data accurately reflects the underlying chemical information, various pre-processing techniques have been developed. The methods described below, aim to mitigate specific artifacts and enhance the quality of the data for subsequent analysis.

- **Centering:** There are various centering algorithms available; however, the most commonly used in spectroscopy is mean-centering. Mean-centering involves calculating the mean for each column of the data matrix and subtracting these mean values from every data point in each spectrum. This approach is particularly useful for removing offsets and enhancing numerical stability during model development [285].

- **Scaling:** These techniques are used to normalize the magnitude of the data. Some methods include auto-scaling, where data is mean-centered and divided by the standard deviation in a column-wise manner. However this approach is rarely used in spectroscopy since it can amplify the influence of noise. Another technique is Min-Max scaling, which transforms the data into a specific range, usually $[0, 1]$, by applying the formula [286]:

$$z_{\text{scaled}} = \frac{z - z_{\text{min}}}{z_{\text{max}} - z_{\text{min}}} \tag{4.12}$$

- **Scattering correction:** Scattering affects the measured spectra, when sample particles have at least one dimension that is roughly the same magnitude as the radiation wavelengths [287]. There are two pre-processing techniques that are widely used; Standard Normal Variate (SNV) and Multiplicative Scatter Correction (MSC). SNV centers and scales each spectrum by subtracting its mean and dividing by its standard deviation, effectively removing baseline shifts and scaling issues. MSC, on the other hand, normalizes spectra by regressing them against a reference spectrum. Both methods are widely used to correct additive and multiplicative effects, improving data quality and enhancing the interpretability of models [288, 289].

- **Spectral Derivation:** Derivation is a powerful technique for correcting spectral distortions, and increasing spectral resolution by highlighting overlapping peaks. The most widely used algorithm is Savitzky-Golay (SG) [290]. SG smooths data by fitting a polynomial to each data point and its N neighbors, then, it calculates the derivative of the fitted polynomial. In broad terms, the first derivative removes additive effects, while the second derivative suppresses both additive effects and baseline slopes. A combination of SNV or MSC with second derivatives can be effective for correcting slopes and multiplicative effects [291].

- **Smoothing and Noise Reduction:** These techniques are used to reduce random noise in the signal. Two common algorithms are wavelet transforms and the use of the SG without the derivation step [286, 290].

- **Wavelength Shift Correction:** Wavelength shifts occur when the "true" spectrum $\mathbf{z}_{\text{true}}$ is shifted relative to the measured spectrum $\mathbf{z}$ by $\delta(j)$ units:

$$\mathbf{z}(j + \delta(j)) = \mathbf{z}_{\text{true}}(j). \tag{4.13}$$

  Techniques for correcting shifts include aligning spectra by maximizing the correlation between measured data and a reference. In NMR spectroscopy, data binning is frequently used, dividing spectra into discrete intervals, integrating the signal, and using these as model inputs [287].

- **Baseline Correction:** Baseline correction aims to remove broad, undesired variations in spectral intensity. One simple approach is polynomial fitting, where a polynomial of order n is fitted to the baseline and subtracted from the spectrum. Advanced methods, such as Asymmetric Least Squares (ALS) [292] and wavelet-based algorithms, offer flexibility in handling more complex baseline shapes. Software packages like R's `baseline` implement these and other methods [286].

Selecting the optimal pre-processing technique is a complex task, often requiring the combination of multiple methods to address the diverse disturbances present in observed spectra. An inappropriate choice of pre-processing can fail to isolate the chemical signal or, even worse, introduce artifacts or eliminate relevant information [287, 293]. The selection process often depends on the characteristics of the data set and the specific requirements of the analysis. As there is no universal pre-processing method suitable for all cases, careful validation is crucial to ensure meaningful and reliable results.

4.3.2 Dimensionality reduction

Many of the experimental techniques used nowadays produce huge amounts of data, difficult to process due to their high dimensionality. In the case of spectroscopic data, a spectrum can contain several thousands of dimensions, which correspond to the acquisition channels. In order to analyze these complex data, several techniques for reducing its dimensionality have been developed, with Principal Component Analysis (PCA) being the most used method.

PCA is a linear transformation technique that projects data onto a new set of orthogonal variables, known as Principal Components (PCs) which are ranked according to the amount of variance they capture in the data [294, Chapter 19]. Principal Components are linear combinations of the original variables. PCA starts by calculating the covariance matrix $\mathbf{C}$:

$$\mathbf{C} = \mathbf{X}^{\mathsf{T}}\mathbf{X}, \tag{4.14}$$

where $\mathbf{X}$ is a mean-centered matrix with dimensions, $N \times M$, in the case of spectral data, the matrix consists of N samples and P wave-channels. Then, PCA finds the direction of the data space with the greatest covariance, this direction can be represented as the vector $\mathbf{v}_{\max}$:

$$\mathbf{v}_{\max} = \arg\max_{\|\mathbf{v}\|=1} \left(\mathbf{v}^{\mathsf{T}}\mathbf{C}\mathbf{v}\right). \tag{4.15}$$

After obtaining $\mathbf{v}_{\max}$, this vector can be used to calculate the eigenvalue λ, which represents the amount of variance in the direction of $\mathbf{v}_{\max}$:

$$\mathbf{v}_{\max} = \mathbf{w} \tag{4.16}$$

$$\lambda = \frac{\mathbf{w}^{\mathsf{T}}\mathbf{C}\mathbf{w}}{\mathbf{w}^{\mathsf{T}}\mathbf{w}}. \tag{4.17}$$

In order to obtain all the eigenvectors of $\mathbf{C}$, all the $\mathbf{w}$ vectors can be arranged as columns of the $\mathbf{W}$ matrix:

$$\mathbf{\Lambda} = (\mathbf{W}^{\mathsf{T}}\mathbf{W})^{-1}(\mathbf{W}^{\mathsf{T}}\mathbf{C}\mathbf{W}) \tag{4.18}$$

$$\mathbf{\Lambda} = \mathbf{W}^{-1}\mathbf{W}^{-\mathsf{T}}\mathbf{W}^{\mathsf{T}}\mathbf{C}\mathbf{W} \tag{4.19}$$

$$\mathbf{\Lambda} = \mathbf{W}^{-1}\mathbf{C}\mathbf{W} \tag{4.20}$$

$$\mathbf{W}\mathbf{\Lambda} = \mathbf{C}\mathbf{W} \tag{4.21}$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues in its elements. From equation (4.21), it is possible to decompose $\mathbf{X}$ into a set of "scores" $\mathbf{T}$ and "loadings" $\mathbf{P}$ (which are the same as the eigenvectors $\mathbf{W}$):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathsf{T}} + \mathbf{E} \tag{4.22}$$

where $\mathbf{E}$ represents the unexplained variance matrix. Although the corresponding eigenvalues and eigenvectors do not have an inherent order, they are typically arranged in decreasing

order of eigenvalue magnitude. To reduce data dimensionality, only a subset of the Principal Components (PCs) is retained, which is related to the signal-to-noise ratio observed in the dataset. Additionally, alternative approaches like cross-validation can guide the selection of the optimal number of PCs [209, 295].

One of the most important uses of PCA is visualization and interpretation. For instance, plots of the data points projected onto the new orthogonal space, known as "scores plots", help to find patterns in the data that may be useful for other purposes, such as classification or outlier detection [209]. Loading/loading plots allow to identify the variable contribution for each PC. In the case of spectroscopic data, it is possible to plot the wavelengths vs the loadings, allowing to show the relevant spectral regions for a given PC.

PCA can also be used to initialize other algorithms. For instance, in Multivariate Curve Resolution (MCR), the number of components in a mixture can be estimated with PCA. Furthermore, more complex dimensionality reduction algorithms, such as t-distributed Stochastic Neighbor Embedding (t-SNE) [210] and Uniform Manifold Approximation and Projection (UMAP) [296], have been shown to better capture the data structure when initialized with PCA [297].

Although PCA is by far the most widely used dimensionality reduction technique, there are scenarios where its application is discouraged. For instance, PCA operates on the assumption that directions of maximum variability are the most significant. This can result in the failure to capture local data structures if the corresponding samples exhibit low covariance relative to the entire dataset. Additionally, in cases of noisy data, PCA may inadvertently prioritize noise in the first Principal Components, compromising the representation of meaningful patterns [297]. On the other hand, advanced non-linear dimensionality reduction algorithms, such as the previously mentioned t-SNE and UMAP, have been developed. These non-linear techniques offer a balance between preserving local and global data structures, making them particularly effective for classification tasks [298]. However, the non-linear transformations involved in these methods complicate the interpretation of the resulting variables, as they lack a straightforward relationship to the original data dimensions

4.3.3 Multivariate regression

Similarly to QSPR, chemometric approaches frequently involve regression tasks to predict target properties based on measured data. In spectroscopy, data obtained from techniques such as MIR, NIR, and Raman spectroscopy are often correlated with reference values. These reference values, obtained using standardized methods, are highly accurate but typically expensive, time-intensive, or require extensive sample preparation [299]. Chemometric regression methods provide an efficient alternative, enabling rapid and minimally invasive measurements with little to no sample preparation. However, these models face challenges such as limited transferability

between different sample matrices or measurement conditions, often necessitating recalibration or adaptation to maintain predictive accuracy [300].

One of the most widely used regression algorithms in chemometrics is Partial Least Squares Regression (PLSR), valued for its robustness in handling highly correlated and noisy data, such as spectroscopic datasets. PLS constructs a set of "latent variables", which are a linear combination of the original variables. These latent variables are designed to maximize the covariance between the predictor matrix $\mathbf{X}$ and the response matrix $\mathbf{Y}$. By capturing the most relevant shared variance, PLS facilitates accurate predictions of Y while mitigating the influence of noise and multicollinearity [301, 302]. Furthermore, as a linear method, PLS offers high interpretability by directly linking predictor variables to the response variable through a linear relationship.

PLS has been extensively combined with various spectroscopic techniques to predict oxidation stability and other properties of oils, diesel, biodiesel, and their blends. For instance, Wen et al. [268] utilized MIR spectroscopy to model the IP of walnut oil samples, while Velvarská et al. [263] employed NIRS and PLS to predict the PetroOxy IP of diesel and biodiesel mixtures. Similarly, Cayuela Sánchez et al. [303] used NIRS spectroscopy to analyze olive oils, predicting the Rancimat IP and other quality parameters such as free acidity, peroxide value, and conjugated dienes.

More recently, regression tasks in chemometrics have adopted more sophisticated algorithms borrowed from machine learning, including SVM, regression trees, and Artificial Neural Network (ANN). These methods offer the ability to model complex, non-linear relationships, often outperforming linear approaches while sacrificing interpretability [304].

A key distinction between our work and the studies mentioned in the literature lies in the focus of the modeling approach. The prior works [263, 268, 303] rely on Beer's law, as they deal with fuel blends where the Induction Period is expressed as a function of the concentrations of mixture components. Conversely, our work focuses on pure hydrocarbons, aiming to represent oxidation stability as a function of molecular structure. This structure is encoded through the presence or absence of specific absorption bands in the spectra.

In this study, we selected SVR as the regression algorithm for its ability to model non-linear relationships and its efficiency in handling datasets with thousands of features, such as spectroscopic data. Compared to methods like XGBoost, SVR offers significantly shorter training times, making it particularly well-suited for this application.

4.4 Materials and methods

4.4.1 Near-Infrared spectroscopy

NIR spectra were acquired using an ABB NIR spectrometer (Model MB3600) equipped with a Deuterated Triglycine Sulfate (DTGS) detector operating in transmission mode. For this study, we only considered the fresh hydrocarbons prior to conducting the accelerated oxidation experiments. The NIR measurements were performed using a 2 ± 0.02 mm cell (QX quality, Hellma) under controlled conditions at a temperature of 27.5 °C, maintained using a Peltier cell. The spectral range covered wavelengths from 833 to 2500 nm (12 000 to 4 000 cm^{-1}). For each sample, 100 scans were conducted to improve the signal-to-noise ratio, and the resulting average spectrum was used for subsequent analysis. The spectrometer was calibrated before measurements to ensure accuracy and reproducibility.

4.4.2 Model development

For the development of the models we employed a framework similar to that used for QSPR-based models. As such, only a brief description of the methodology will be provided here.

The NIR spectra ranging from 833 to 2265 nm for 82 pure hydrocarbons were utilized, resulting in a matrix of dimensions 82×3933. This matrix served as the reference for generating the regression matrix. To achieve this, we considered the number of unique temperature-sample combinations. Each hydrocarbon's spectrum was replicated based on these combinations, with the experimental temperature (in Kelvin) appended as an additional column (figure 4.6). Consequently, the final matrix had dimensions of 205×3934.

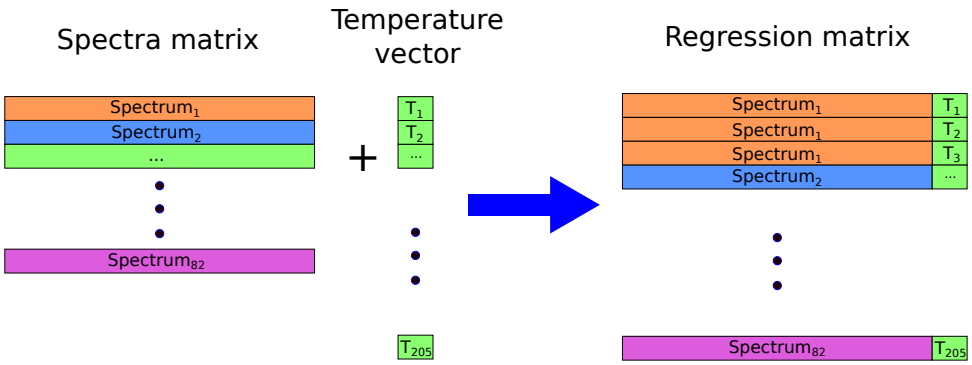


FIGURE 4.6 – Generation of the regression matrix for the NIR-based model.

In this work, we applied various pre-processing algorithms to treat the spectra before proceeding to model development. Specifically, we employed several methods, including Iteratively Reweighted Least Squares (IRLS) [305, Chapter 7] for baseline correction, MinMax scaling, and Savitzky-Golay for spectral differentiation. Additionally, we tested multiple

combinations of these methods to determine their impact on the resulting models. For the modeling step, we exclusively used SVR-RBF. This choice was made because XGBoost, in both its linear and regression tree implementations, proved to be too computationally intensive given the high dimensionality of the data. We emphasize the need of a non-linear algorithm to account for the strong non-linear relationships between molecular structure features and the induction period, as previously discussed in Chapter 2. Due to the positive data skewness, we performed the modeling on the log-transformed IP value. As in the previous section, all data manipulation and modeling was performed using the R programming language [231], along the `caret` [232], `kernlab` [233] and packages. We also used the `baseline` [306] and `hyperspec` [307] packages for the implementation of the IRLS algorithm and the manipulation of spectra data files, respectively.

Model validation followed a nested cross-validation approach as previously described. This approach included 10 internal folds ($i = 10$) for hyper-parameter tuning and 5 external folds ($k = 5$) for model evaluation. To ensure diversity in the training set, the Interesting Feature Finder algorithm [238] was employed to identify the most distinct spectra in the dataset. Using this algorithm, five extreme spectra; cyclopentane, n-propylbenzene, n-butylcyclohexane, and 1-hexene, were selected and fixed in the training set. Database splitting was performed randomly using a compound-out strategy. Under this approach, when a molecule was assigned to a specific fold, all its instances (IP measurements at various temperatures) were included in the same fold. As in the previous case, we aimed to minimize the RMSE during model optimization.

4.5 Results

In this section, the discussion of the NIR spectra is divided into two parts. The first part focuses on the analysis of the NIR spectra and the assignment of spectral bands, which is essential for interpreting the NIR-based models. The second part discusses the results of the modeling step, including the effects of various pre-processing techniques on accuracy and a comparison with our best QSPR-based model.

4.5.1 Spectral analysis

4.5.1.1 Band assignment

In this study, we only measured the NIR spectra for the liquid samples in our database, resulting in a total of 84 samples. Figure 4.7 shows the raw NIR spectra, colored by hydrocarbon family. From the spectra, it is evident that the signal in the 2265–2500 nm region is saturated, requiring its removal before further analysis. Additionally, two samples, corresponding to 1,5-hexadiene

and 1,7-octadiene, exhibit a saturated band at 2230 nm. As this issue is exclusive to these two samples, we opted to exclude them from our dataset to prevent further loss of spectral information. Thus, the NIR spectra, from 833 to 2265 nm, for the remaining 82 samples, are presented in figure 4.8.

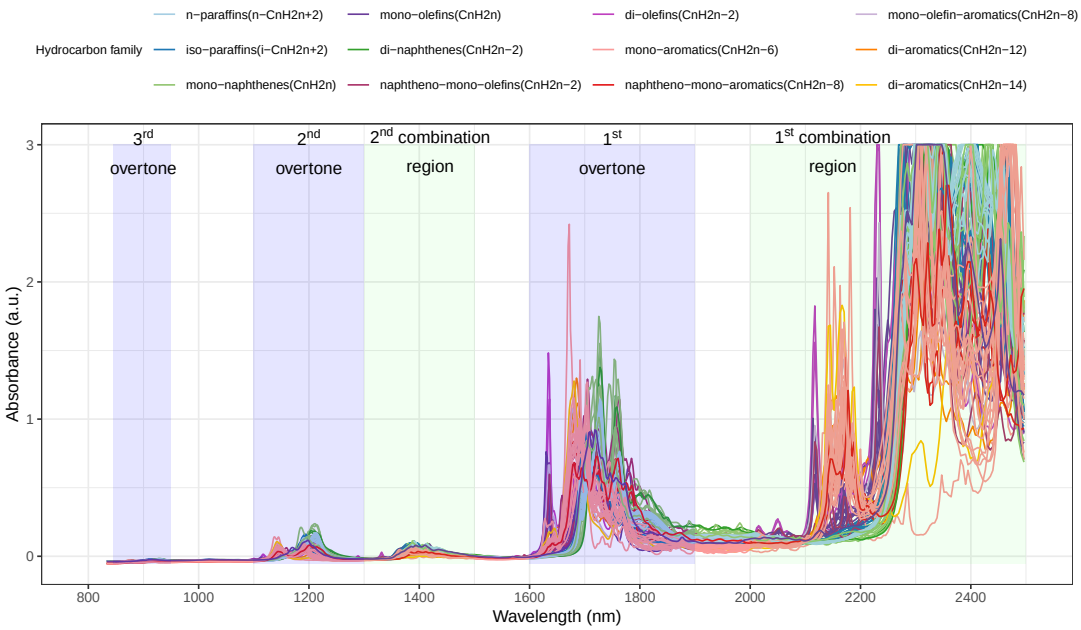


FIGURE 4.7 – NIR spectra from 833 to 2500 nm, for the 84 compounds analyzed in this study. Overtone regions are highlighted in blue and the combination regions in green.

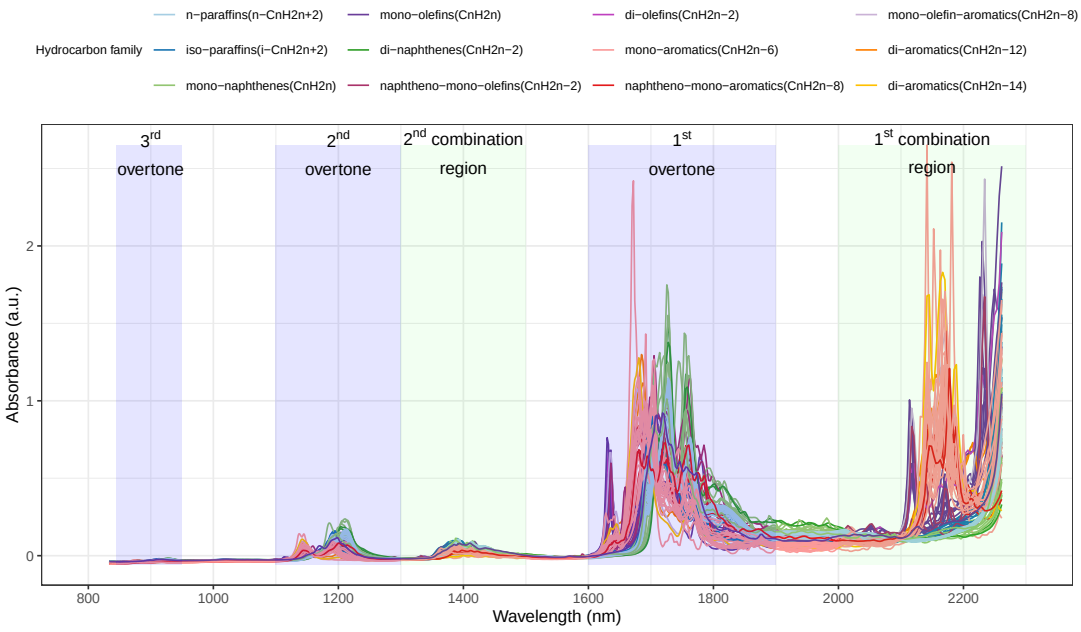


FIGURE 4.8 – Truncated NIR spectra from 833 to 2265 nm for the remaining 82 compounds. Overtone regions are highlighted in blue and the combination regions in green.

In the following analysis, we performed spectral band assignment based on the data reported by Workman and Weyer [259, Chapters 2-4].

First, we examine the NIR spectra of linear alkanes. For this purpose, we selected *n*-pentane, *n*-decane, and *n*-hexadecane, representing the shortest, an intermediate, and the longest liquid linear alkanes in our database, respectively. As shown in figure 4.9, several characteristic bands can be identified:

1. **First Combination Region:**
   - The methyl group (CH_3) exhibits a combination of asymmetric stretching and bending vibrations ($\upsilon_a + \delta$).
2. **First Overtone Region:**
   - Four prominent bands are observed, two for methyl (CH_3) and two for methylene (CH_2) groups:
     - Methyl groups: A band at 1693 nm and a second band at $(1709 - 12.5 \cdot F)$ nm, where F is the CH_3 mole fraction in a given hydrocarbon.
     - Methylene groups: A band at 1763 nm and an asymmetric/symmetric stretching combination ($\upsilon_a + \upsilon_s$) at $(1708 + 25.1 \cdot W)$ nm, where W is the CH_2 weight fraction.
3. **Second Combination Region:**
   - Signals arise from the combination of the first overtone and bending vibrations ($2\upsilon + \delta$) of methyl and methylene groups:
     - Methylene groups: A double peak at 1392 and 1412 nm.
     - Methyl groups: A double peak at 1360 and 1377 nm.
4. **Second Overtone Region:**
   - Strong bands are observed for methyl and methylene groups at 1192 and 1210 nm, respectively.
   - Additional weaker methyl bands appear at 1153 and 1176 nm.
5. **Third Overtone Region:**
   - Methyl and methylene groups exhibit bands at 913 and 929 nm, respectively.

Overall, the NIR spectra show that methyl (CH_3) signals are more intense in short-chain linear paraffins due to their higher CH_3:CH_2 ratio compared to long-chain linear paraffins.

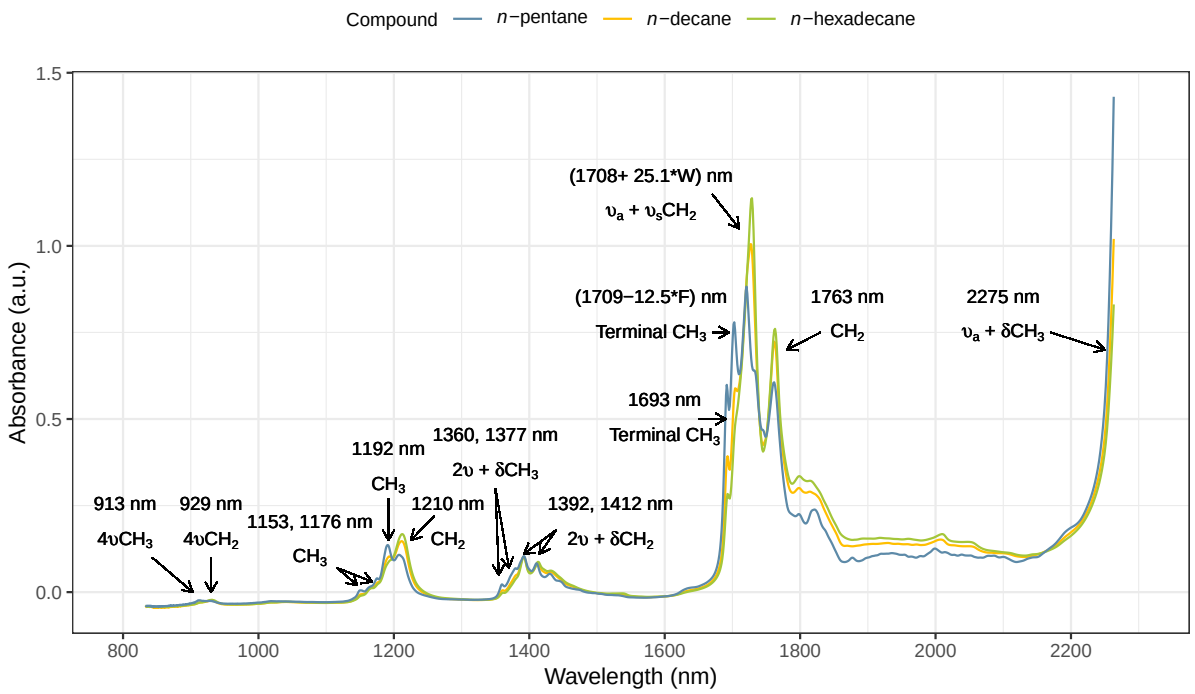


FIGURE 4.9 – **a**) NIR spectra of linear alkanes, including *n*-pentane, *n*-decane and *n*-hexadecane, from 833 to 2265 nm. Band assignments are highlighted.

In the case of branched alkanes, we present the spectra of several C_8 isomers, namely 2-methylpentane, 2,3,4-trimethylpentane, and 2,2,4-trimethylpentane. For comparison, the spectrum of *n*-octane is also included. Since the spectral band assignments are similar to those of linear alkanes, we focus on discussing the differences.

As shown in figure 4.10**a**, the intensity of the CH_3-related bands increases with the degree of substitution. For example, the band at 1693 nm exhibits an absorbance of approximately 0.8 for 2,3,4-trimethylpentane, compared to only 0.4 for *n*-octane.

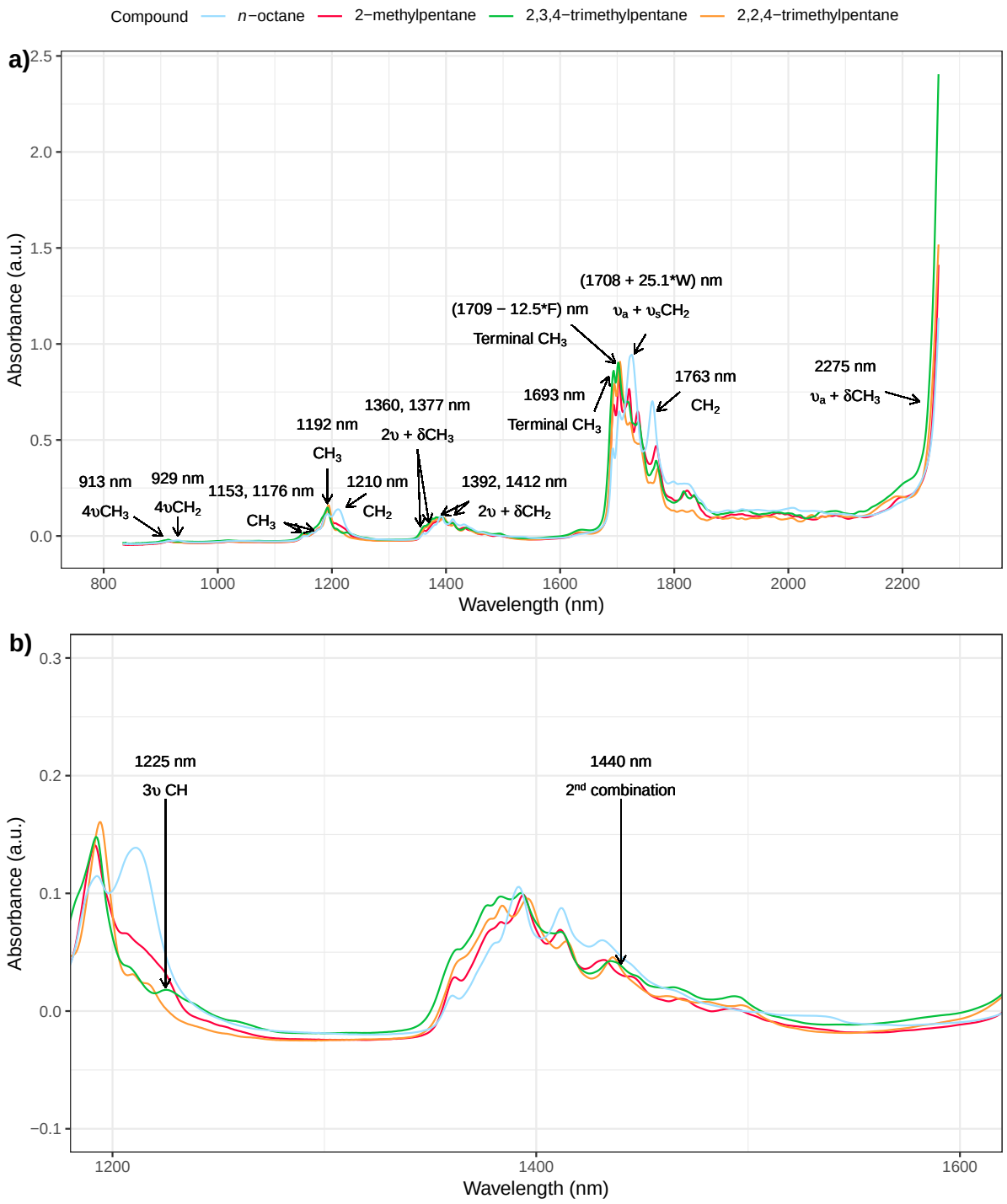


FIGURE 4.10 – **a)** NIR spectra of branched alkanes, including 2-methylpentane, 2,3,4-trimethylpentane, and 2,2,4-trimethylpentane, from 833 to 2265 nm. The spectrum of *n*-octane is included for comparison. Band assignments are highlighted. **b)** A magnified view of the NIR spectra for the same compounds, focusing on the 1190 to 1610 nm region to provide a detailed examination of methine (CH) spectral bands.

In contrast, signals associated with the methine group (CH) are significantly weaker than those of methyl and methylene groups. For instance, the second first overtone (2υ) of the methine group is expected to appear in the 1700–1800 nm region. However, Wheeler [308]

reported the presence of the 2nd overtone (3υ) and the 2nd combination bands at 1225 nm and 1440 nm, respectively. In figure 4.10**b**, the 3υ signal for 2,3,4-trimethylpentane, a compound with three methine groups, can be observed. Nevertheless, this signal overlaps with the methyl group absorption at 1210 nm, as shown for 2-methylpentane. Similarly, the 2nd combination band of methine groups is also absorbed, however other compounds without substitutions also exhibit absorption in the same region. Lastly, quaternary carbon atoms cannot be detected with this technique due to their lack of C−H bonds. Consequently, the weak absorption of methine groups and the absence of signals from quaternary carbons present a significant limitation for predicting the Induction Period from NIR spectral data.

For cycloalkanes, the NIR spectra are presented in figure 4.11. The first combination band is partially observed at 2220 nm and is hypothesized to arise from stretching and bending vibrations ($\upsilon + \delta$). In the first overtone region, two distinct bands at 1727 nm and 1755 nm can be identified, corresponding to asymmetric ($2\upsilon_a$) and symmetric ($2\upsilon_s$) stretching vibrations, respectively.

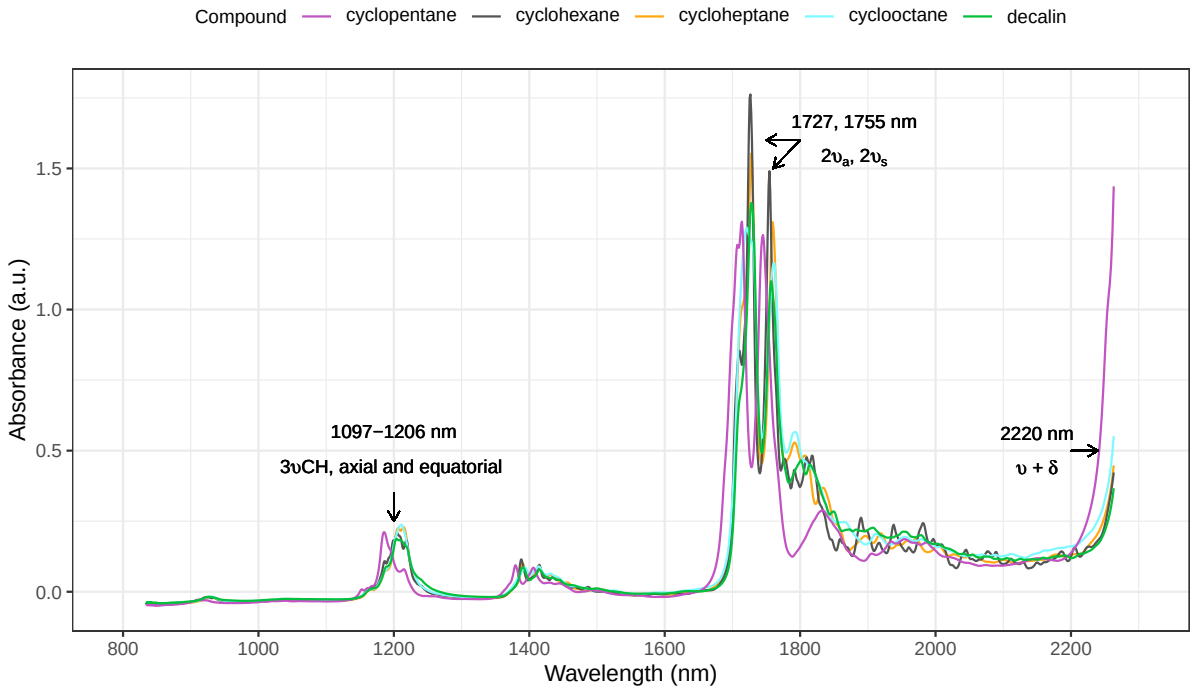


FIGURE 4.11 – NIR spectra of cycloalkanes, including cyclopentane, cyclohexane, cycloheptane, cyclooctane and decalin, from 833 to 2265 nm. Band assignments are highlighted.

In the 2nd overtone region, a band associated with C−H stretching from carbons in the axial and equatorial positions of the ring is observed. Ring strain significantly influences the spectral features of cycloalkanes. For example, in strained molecules such as cyclopentane, the 3υCH band appears at approximately 1180 nm, whereas in less-strained cycloalkanes like cyclohexane, cyclooctane, and decalin, this band is observed at around 1206 nm. Additionally,

cyclopentane lacks the unassigned band at 1790 nm, which is present in the spectra of the other cycloalkanes.

The NIR spectra of various alkenes are shown in figure 4.12. Below, the corresponding band assignments are detailed.

- **First Combination Region:**
  - Four bands are observed:
    - **Three bands** at approximately 2120, 2232, and 2174 nm are associated with bending vibrations of the CH_2 group in the vinyl group ($CH_2{=}CH^-$) of 1-alkenes.
    - **Fourth band** at 2169 nm corresponds to the asymmetric stretching ($2\upsilon_a$) of the terminal allylic CH_2, observed in *trans*-2-hexene.
- **First Overtone Region:**
  - Two bands are observed:
    - The first band corresponds to the first overtone of the C=C bond, appearing at approximately 1677 nm.
    - The second band is related to the stretching vibration of allylic C−H, appearing between 1613 and 1639 nm. This band is absent in *trans*-2-hexene and 2,3-dimethyl-2-butene.
- **Second Combination Region:**
  - Three bands are identified:
    - Two bands at 1290 and 1361 nm are high-order overtones of the 2120 and 2232 nm bands.
    - The third band at 1332 nm is the second member of the 2174 nm band progression.
- **Second Overtone Region:**
  - A band between 1118 and 1124 nm corresponds to the $3\upsilon CH_2$ vibration.
- **Third Overtone Region:**
  - The 3rd and 4rd overtones are observed at approximately 878 nm and 873 nm, respectively.

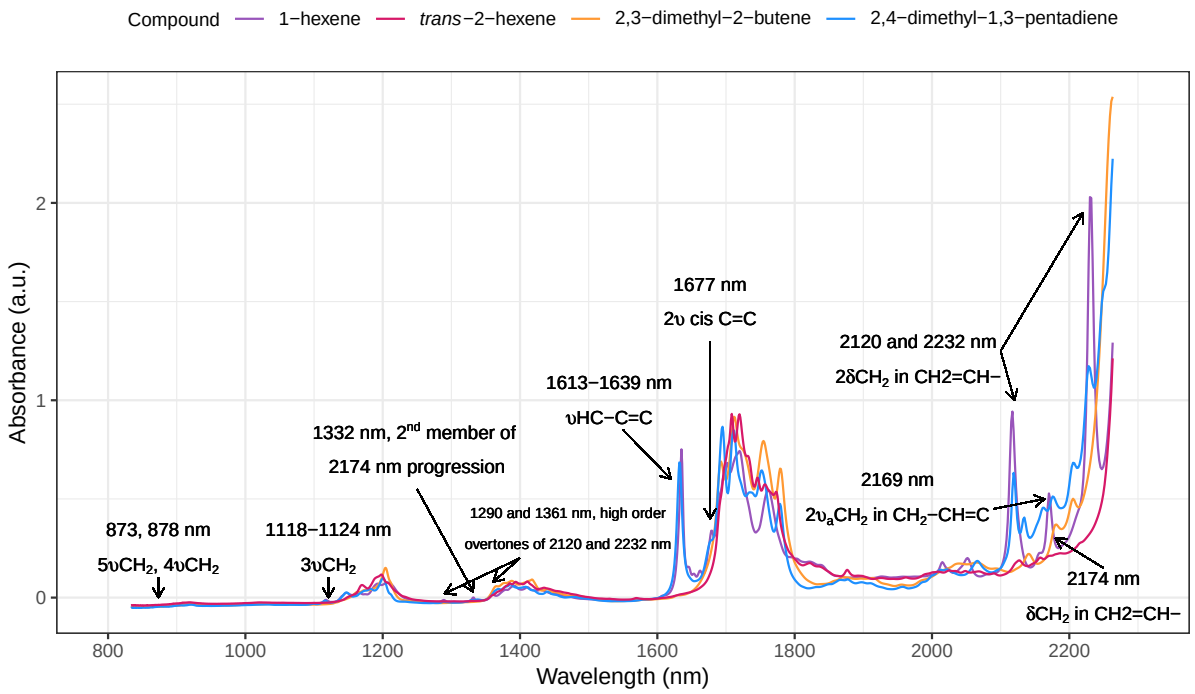


FIGURE 4.12 – NIR spectra of cycloalkanes, including 1-hexene, *trans*-2-hexene, 2,3-dimethyl-2-butene, and 2,4-dimethyl-1,3-pentadiene, from 833 to 2265 nm. Band assignments are highlighted.

The NIR spectra of cycloalkenes is presented in figure 4.13. In the first combination region, an unassigned band is observed at 2141 nm in cycloalkenes. In the first overtone region, another unassigned band appears at 1667 nm, which is exclusive to cyclopentene. Additionally, a small double peak at 1653 nm is attributed to the stretching vibrations of the allylic CH_2 groups. Lastly, the second overtone of the vinyl group is observed at 1139 nm.

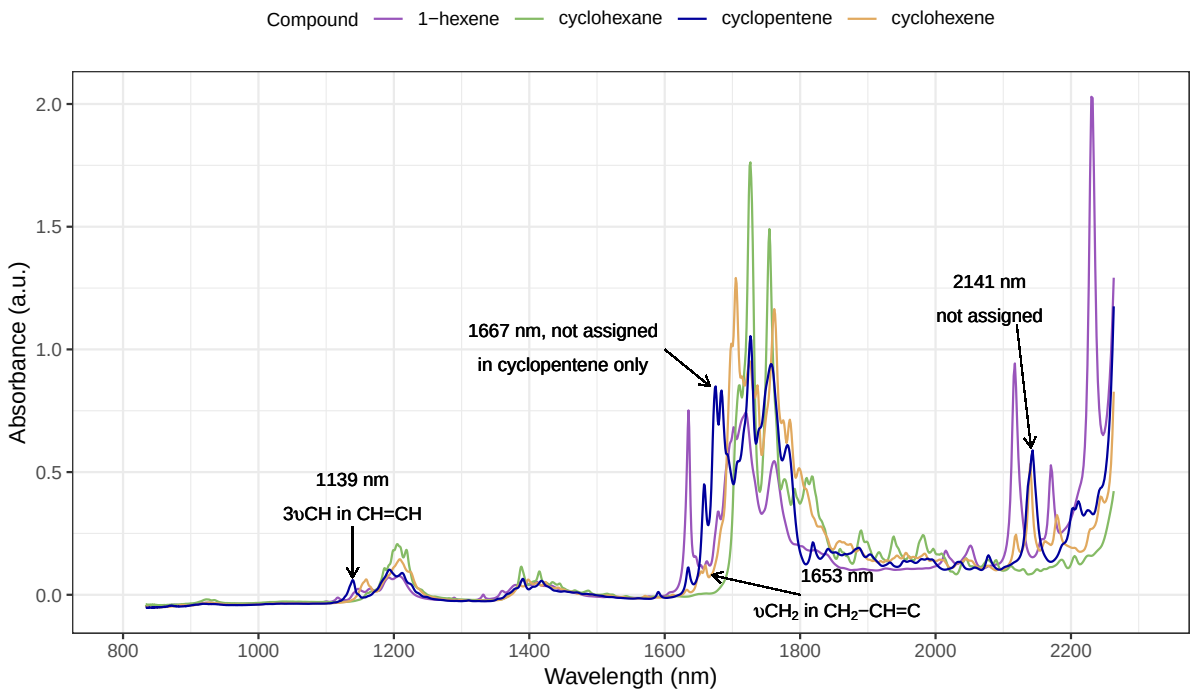


FIGURE 4.13 – NIR spectra of cycloalkenes, including cyclohexene and cyclopentene, from 844 to 2265 nm. The spectra of 1-hexene and cyclohexane are included for comparison. Band assignments are highlighted.

Finally, the NIR spectra of aromatic compounds are shown in figure 4.14. The NIR spectrum of benzene is particularly interesting because its fundamental vibrational bands are absent in the mid-infrared region due to the molecule's symmetry. However, its overtone and combination bands are prominent in the near-infrared region. In the first combination region, five bands can be observed at 2148, 2154, 2167, and 2188 nm. These bands primarily correspond to combinations of CH stretching vibrations. Notably, the intensity of these bands decreases with the presence of substitutions, as seen in 1,2,4-trimethylbenzene, and *n*-octylbenzene. In the first overtone region, two combination bands appear for benzene at 1671 and 1689 nm. Additionally, the first overtone of methyl groups ($2\upsilon CH_3$) is visible at 1767 nm for alkylbenzene derivatives. The second overtone of the aromatic C−H bonds is observed at 1132 nm, while the overtone associated with substituents appears at 1192 nm. Lastly, the third overtone is detected at 874 nm.

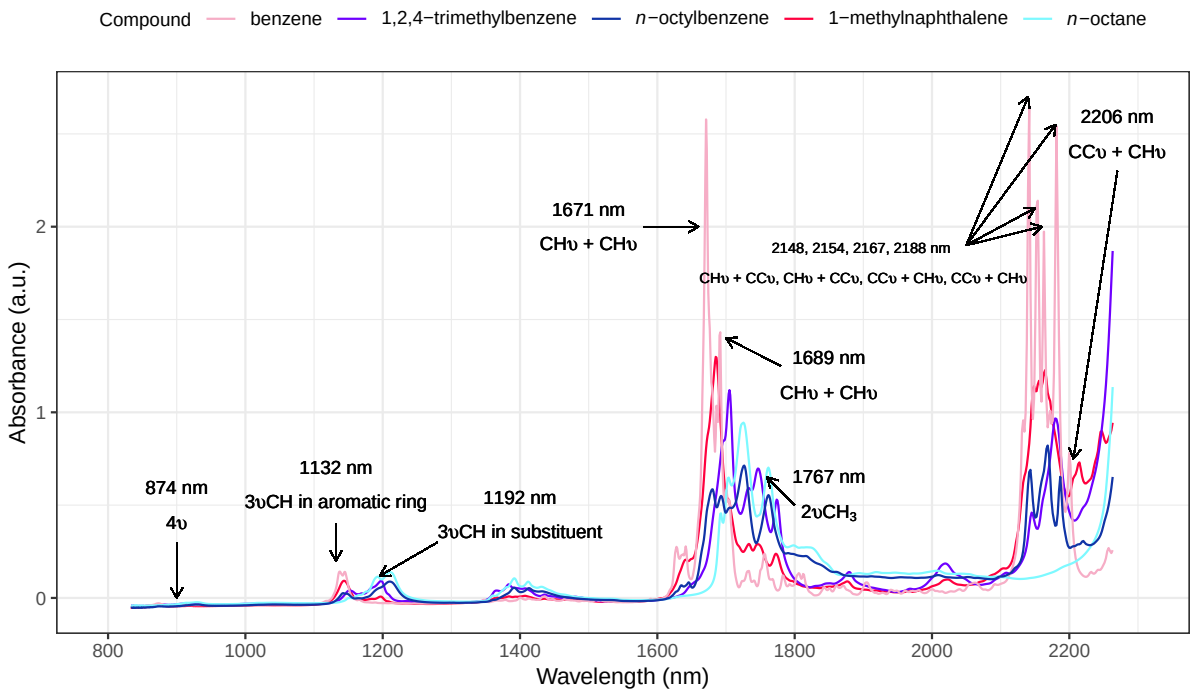


FIGURE 4.14 – NIR spectra of aromatic compounds, including benzene, 1,2,4-trimethylbenzene, and *n*-octylbenzene, from 833 to 2265 nm. The spectrum of *n*-octane is included for comparison. Band assignments are highlighted.

4.5.1.2 Data visualization / Dimension reduction

In this section, we show the results from the PCA performed on the 82 mean-centered NIR spectra, in the spectral range described in section 4.5.1.1. Figure 4.15 shows the cumulative variance explained as a function of the number of principal components. As depicted, the first eight PCs account for approximately 95% of the variance, while 18 PCs explain 99% of the variance.

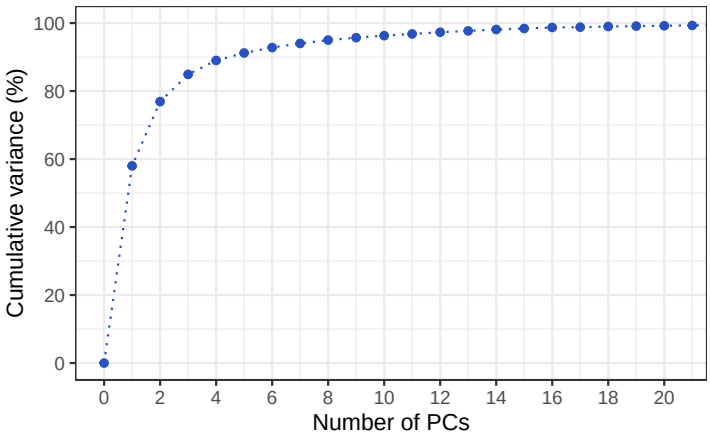


FIGURE 4.15 – Cumulative variance explained by the first 20 Principal Components (PCs) of the dataset. The plot shows that the first few PCs capture the majority of the variance.

We discuss the first three PCs, as they represent approximately 80% of the variance alone. Figure 4.16 shows the scores plot for PC1 vs. PC2. In figure 4.16**a**, the color gradient reflects the log(IP) values, ranging from low (blue) to high (red). It is not possible to observe a clear pattern relating the IP and the NIR spectra based on this projection.

Figure 4.16**b** illustrates the formation of overlapping clusters corresponding to different hydrocarbon families. Saturated compounds, including *n*-paraffins, iso-paraffins, and naphthenes, are positively projected along PC1, while mono- and di-aromatics are negatively projected. Olefins and di-olefins are found near the coordinate 0. This distribution indicates that PC1 effectively discriminates between aromatics, olefins, and saturated compounds.

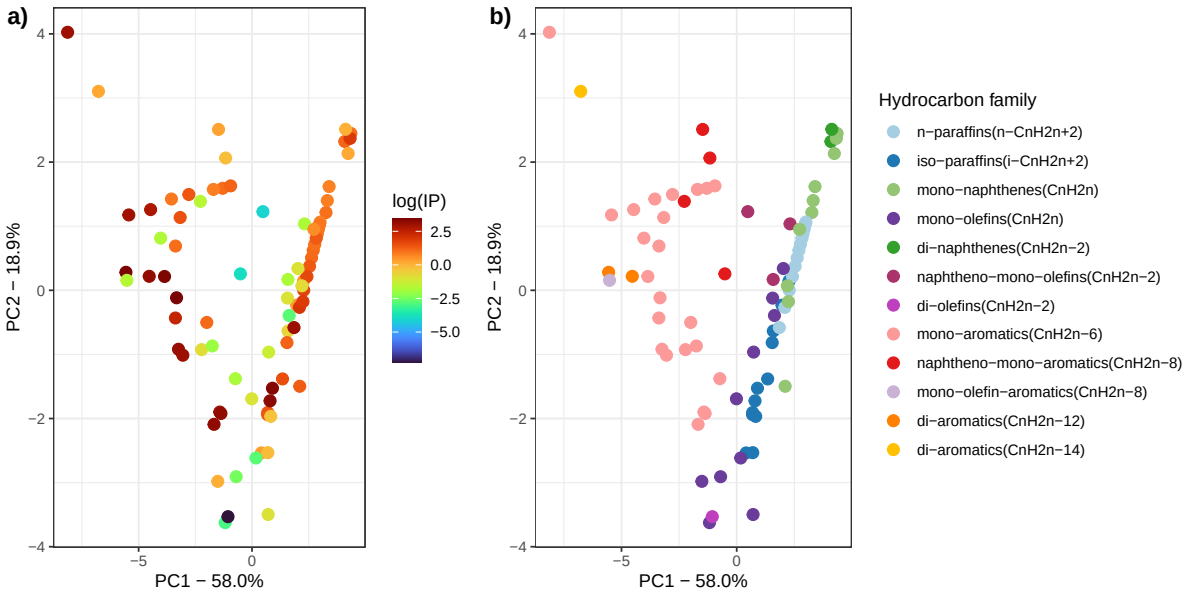


FIGURE 4.16 – Scores plot for PC1 vs. PC2 based on raw NIR spectra. The data points are color-coded to represent **a)** log(IP) values and **b)** hydrocarbon families.

This separation is further corroborated when observing the loadings plot of PC1, and the spectra of the samples projected on the two extremes of PC1 (see figure 4.17). The samples located at the extreme left (benzene) and extreme right (cyclopentane) in the scores plot correspond to distinct spectral features. The negative loadings in PC1 are dominated by bands near 1130 nm, 1690 nm, and 2150–2200 nm. These bands correspond to the second overtone of the C—H bond in aromatic rings, as well as the first overtones and combination bands typical of aromatic compounds (figure 4.14). Conversely, the bands with positive contributions to the loadings are located at 1206, 1727, and 1755 nm. These bands represent the second and first overtones of methylene groups, which are characteristic of saturated compounds (figures 4.9 to 4.11). Interestingly, the aromatic compounds form a horizontal pattern on the plot, spanning from coordinates (-5.2, 1.6) to (-0.5, 1.8). This arrangement corresponds to aromatics with increasing length of attached paraffinic chains, ranging from toluene to *n*-octylbenzene. As the

chain length increases, the samples are projected further to the right, reflecting the increasing paraffinic character of these compounds.

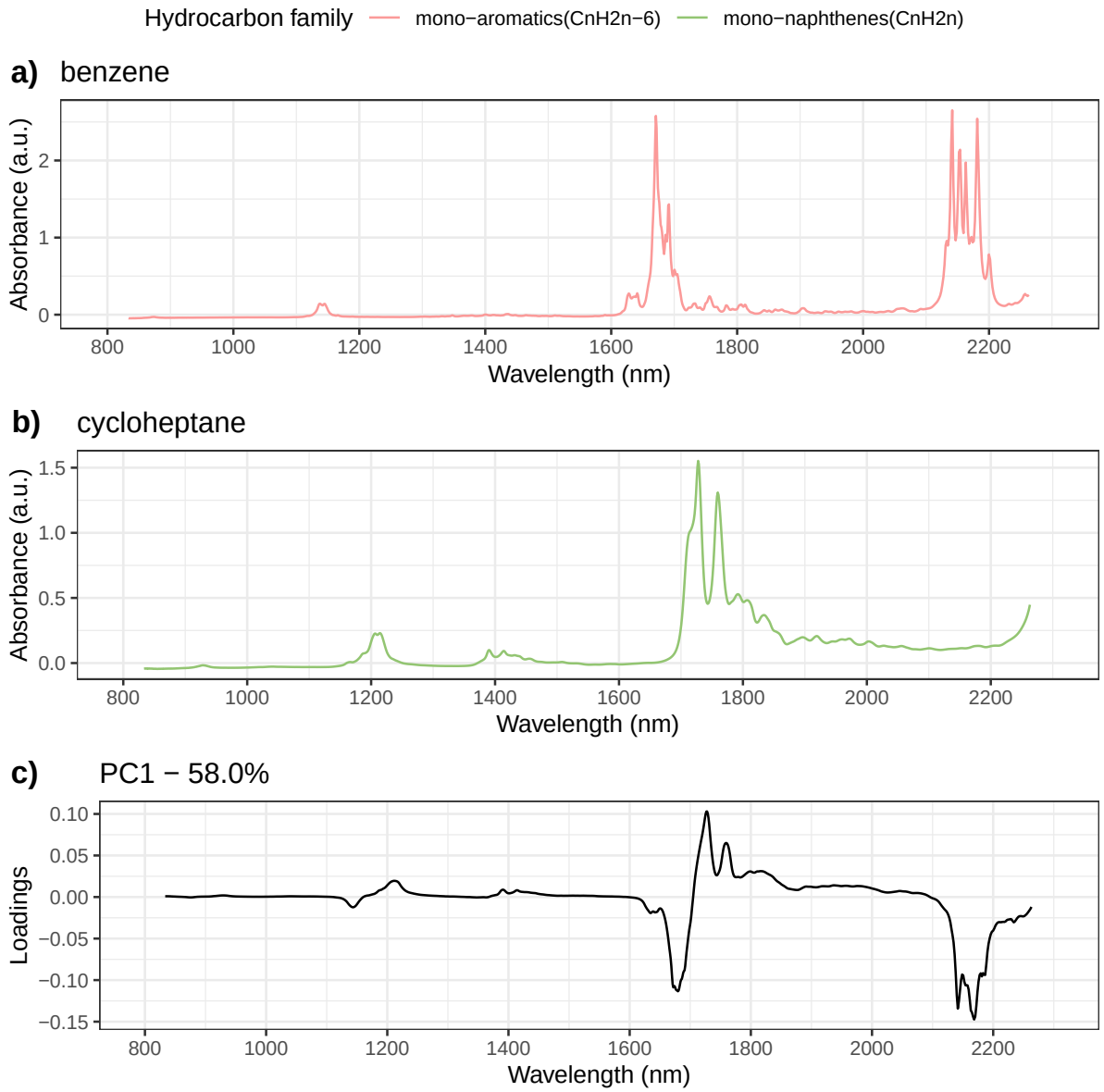


FIGURE 4.17 – Spectral analysis associated with PC1. **a)** Spectrum of benzene, the compound with the lowest PC1 score. **b)** Spectrum of cycloheptane, the compound with the highest PC1 score. **c)** Loadings for PC1, which explains 58.0% of the variance

The interpretation of PC2 in figure 4.16 provides additional insights into the separation of hydrocarbon families. As shown in figure 4.18, the bands contributing negatively to the loadings are associated with combination bands that are present across most hydrocarbon families, such as those in the 1350–1400 nm region. Additionally, contributions from terminal CH_3 groups at 1700 nm and vinyl and allyl groups at 1150–1200 nm, 1350–1400 nm, 1620 nm, 1750 nm, 2120 nm, and >2220 nm are also significant.

On the other hand, the bands contributing positively to the loadings are linked to CH_2 groups, which are observed at 1210 and 1720 nm, as well as to the first overtones of methyl groups at 1680 and 1760 nm. Furthermore, combination bands present in aromatic compounds, specifically in the 2120–2180 nm range, also contribute positively to the loadings.

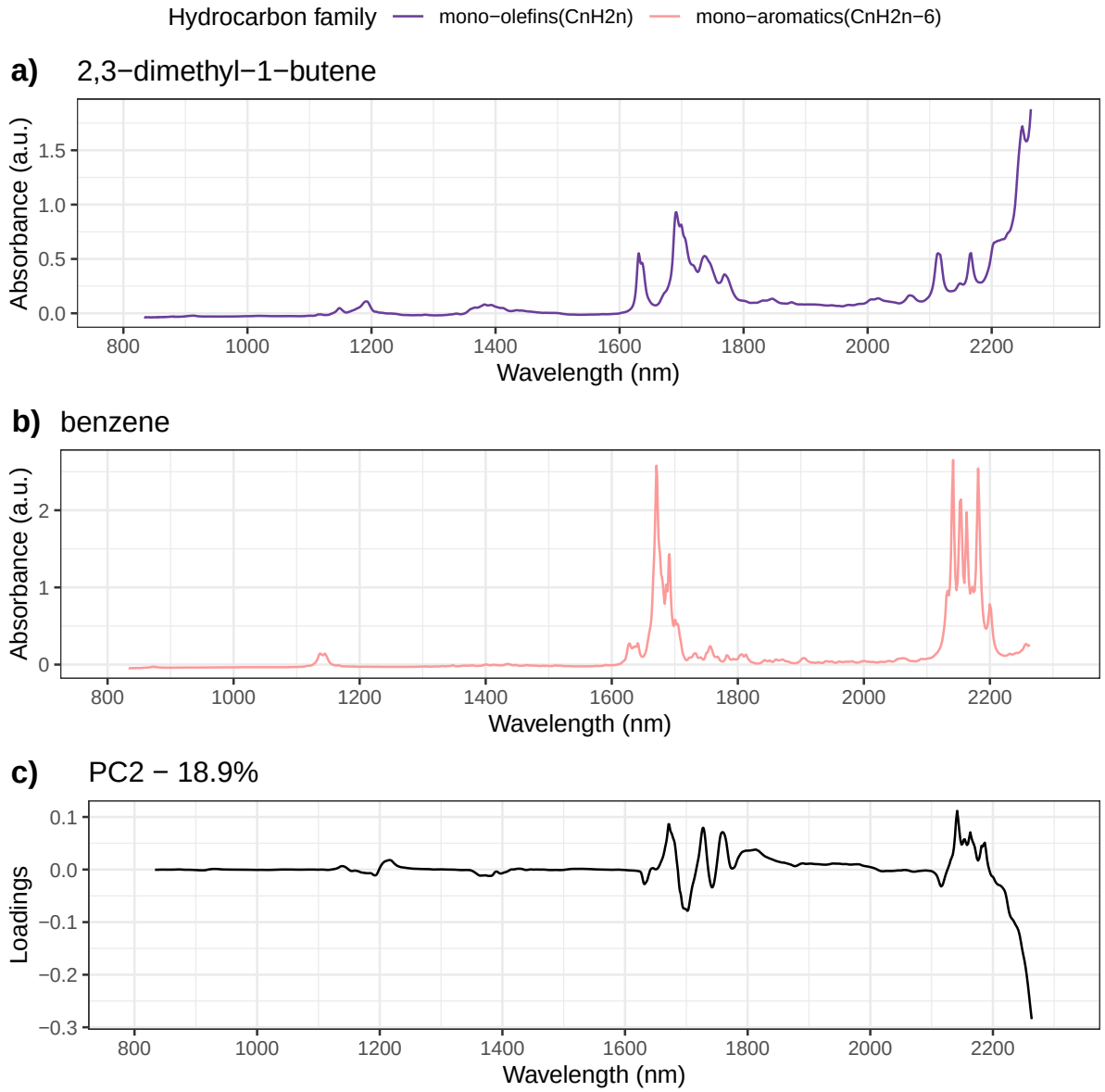


FIGURE 4.18 – Spectral analysis associated with PC2. **a)** Spectrum of 2,3-dimethyl-1-butene, the compound with the lowest PC2 score. **b)** Spectrum of benzene, the compound with the highest PC2 score. **c)** Loadings for PC2, which explains 18.9% of the variance

Samples located in the uppermost region of the plot correspond to aromatic and paraffinic compounds, whereas those in the lowermost region are primarily olefins and highly branched paraffins. Linear alkanes display a clear vertical pattern, with longer-chain compounds positioned toward the top and shorter-chain compounds toward the bottom. Long-chain alkenes are found close to the linear alkanes due to their strong paraffinic character. In contrast,

short-chain alkenes and highly branched alkanes are positioned in the lower region of the plot, reflecting their distinct spectral characteristics.

PC3 is plotted against PC1 and PC2 in figures 4.19 and 4.20. These plots show that PC3 achieves a certain level of discrimination based on IP values, with reactive compounds predominantly located in the upper regions of the plots. This separation is largely attributed to the positioning of olefins in this area, as these compounds are highly reactive.

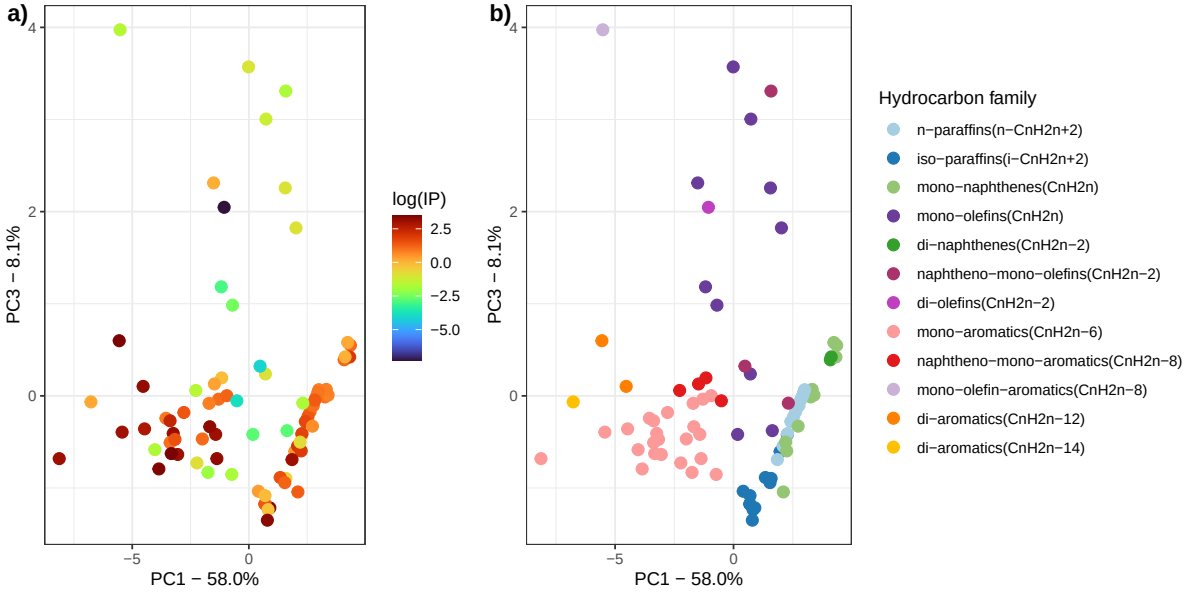


FIGURE 4.19 – Scores plot for PC1 vs. PC3 based on raw NIR spectra. The data points are color-coded to represent **a)** log(IP) values and **b)** hydrocarbon families.

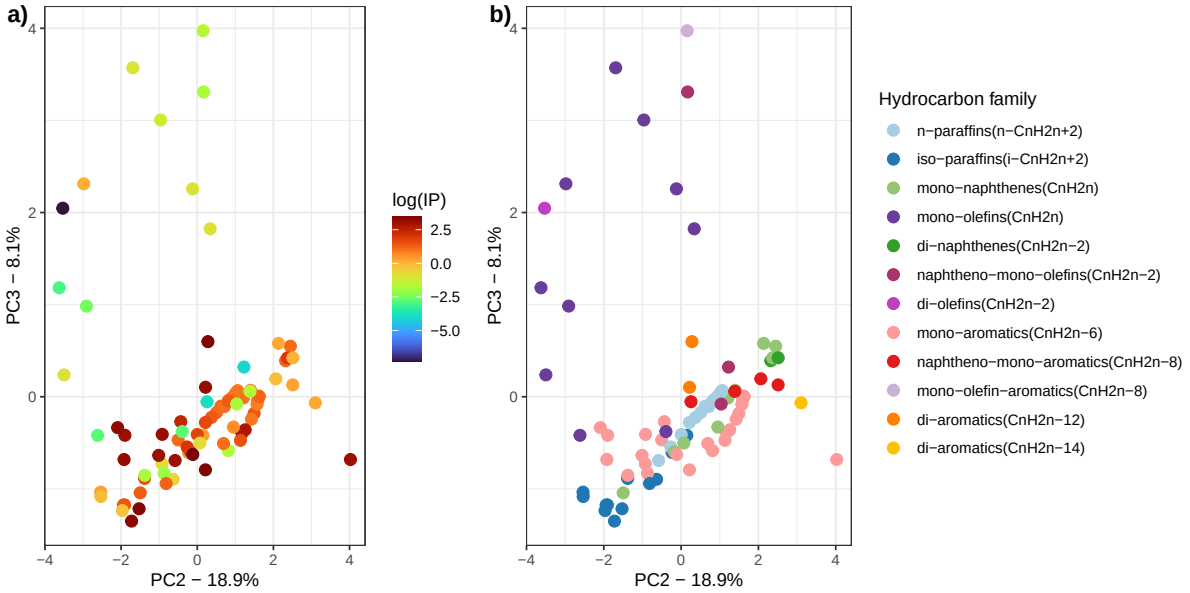


FIGURE 4.20 – Scores plot for PC2 vs. PC3 based on raw NIR spectra. The data points are color-coded to represent **a)** log(IP) values and **b)** hydrocarbon families.

When examining the loadings of PC3, as presented in figure 4.21, it becomes evident that the bands contributing negatively to the loadings are found within specific regions: 1690–1710 nm, associated with the first overtone of methyl groups; 1740 nm, corresponding to the first overtone of methylene groups; and wavelengths exceeding 2250 nm, which correspond to the combination band $\upsilon_a + \delta CH_3$ found in paraffinic compounds. Conversely, bands contributing positively to the loadings are located at 1620 nm, related to the stretching vibration of allylic carbons in 1-alkenes; at 1720 and 1760 nm, associated with methylene groups; and at 2120 and 2230 nm, corresponding to the first overtone of vinylic groups. As a result, olefins are positioned in the upper regions of the plots, while aromatics and paraffins occupy the lower regions.

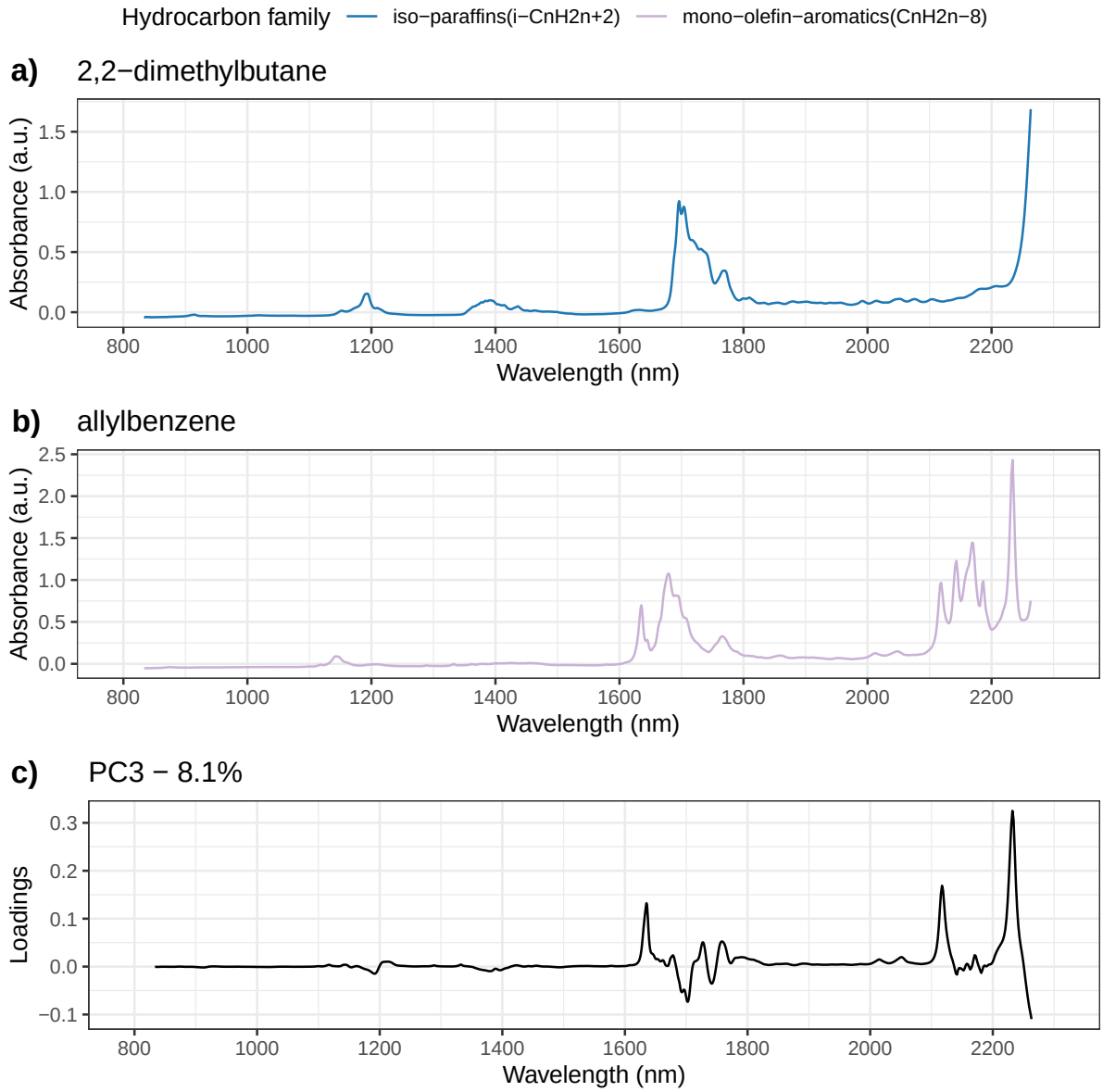


FIGURE 4.21 – Spectral analysis associated with PC2. **a)** Spectrum of 2,2-dimethylbutane, the compound with the lowest PC3 score. **b)** Spectrum of allylbenzene, the compound with the highest PC3 score. **c)** Loadings for PC3, which explains 8.1% of the variance

The scores plots for the first eight PCs, explaining 95% of the variance are provided in Appendix E.

4.5.2 Model performance

This section presents the results from the developed models. Table 4.2 summarizes the optimal parameters and goodness-of-fit metrics for various pre-processing techniques. Among the techniques evaluated, baseline correction combined with MinMax scaling achieved the lowest RMSEP, indicating the best predictive performance. Conversely, the Savitzky-Golay first derivative method demonstrated the poorest performance, with relative prediction errors of 161% and 1200%, respectively.

Table 4.2 – Model performance metrics for SVR-RBF for NIR-based models. RMSE values were averaged across the 5 external cross-validation folds.

Pre-processing	**Optimal hyper-parameters**	**Scale**	**RMSEC**	**RMSECV**	**RMSEP**
None	$\sigma = 0.001$ $C = 1000$ $\epsilon = 0.10$	log hour	0.54 9.82	0.63 11.87	1.17 20.87
Baseline	$\sigma = 0.001$ $C = 1000$ $\epsilon = 0.00$	log hour	0.45 9.69	0.64 11.94	1.16 19.26
MinMax	$\sigma = 0.001$ $C = 1000$ $\epsilon = 0.15$	log hour	0.42 8.98	0.55 10.56	1.25 19.05
SG-1	$\sigma = 0.1$ $C = 1584$ $\epsilon = 0.05$	log hour	0.53 10.5	0.48 9.63	1.52 131.41
SG-2	$\sigma = 0.01$ $C = 63095$ $\epsilon = 0.00$	log hour	0.97 14.76	1.27 19.07	1.46 28.89
Baseline+ MinMax	$\sigma = 0.0006$ $C = 2511$ $\epsilon = 0.10$	hour log	0.39 9.56	0.52 9.92	1.13 17.61
Baseline + MinMax + SG1	$\sigma = 0.016$ $C = 3981$ $\epsilon = 0.10$	log hour	0.75 13.63	0.32 8.46	1.42 22.54
Baseline + MinMax + SG2	$\sigma = 0.03$ $C = 10000$ $\epsilon = 0.10$	log hour	0.96 14.87	1.12 16.33	1.39 21.31

None of the NIR-based models outperformed the non-data-augmented models based on molecular descriptors. For instance, the best-performing QSPR model, XGBoost Linear, achieved a relative prediction error of 123%, while the worst one, obtained with XGBoost Tree, had an error of 148%. For reference, the SVR-RBF model yielded a relative error of 145%.

Given the high error associated with the NIR-based models, parity plots for models employing various pre-processing techniques are presented on a logarithmic scale in figure 4.22. For the best-performing model, shown in figure 4.22**f**, the hydrocarbons with the worst predictions include 2,2-dimethylbutane, 2,2,4-trimethylpentane, 2,3-dimethyl-2-butene, cumene, benzene, and cyclopentane.

The underestimation of the iso-paraffins 2,2-dimethylbutane and 2,2,4-trimethylpentane could be attributed to the presence of quaternary carbons, which are not directly observable in the NIR spectrum. Although the spectra highlight the high number of methyl groups in these compounds, the model appears to associate an abundance of methyl groups with low stability, leading to erroneous predictions. For 2,3-dimethyl-2-butene, the model significantly overestimates its stability. This is likely because the molecule lacks characteristic alkene bands at 1610, 2120, and 2232 nm (see figure 4.12), which are typically associated with the hydrogen atoms in vinyl groups. In this case, the absence of such hydrogen atoms (due to the complete substitution of the vinyl group) may affect the model predictions.

Aromatic compounds such as cumene and benzene also show poor predictions. For cumene, the weak signal from the methine group likely contributes to this issue, as it is barely detectable in its NIR spectrum. Benzene, on the other hand, exhibits intense bands at 2148, 2154, 2167, 2188, and 2206 nm that are not observed in other aromatic compounds (figure 4.14), potentially skewing the model's interpretation. Lastly, cyclopentane may be poorly predicted due to its spectral bands being shifted to lower wavelengths, a consequence of ring strain (figure 4.11). In this case, applying wavelength-shift correction as a pre-processing technique could improve the model's predictive performance.

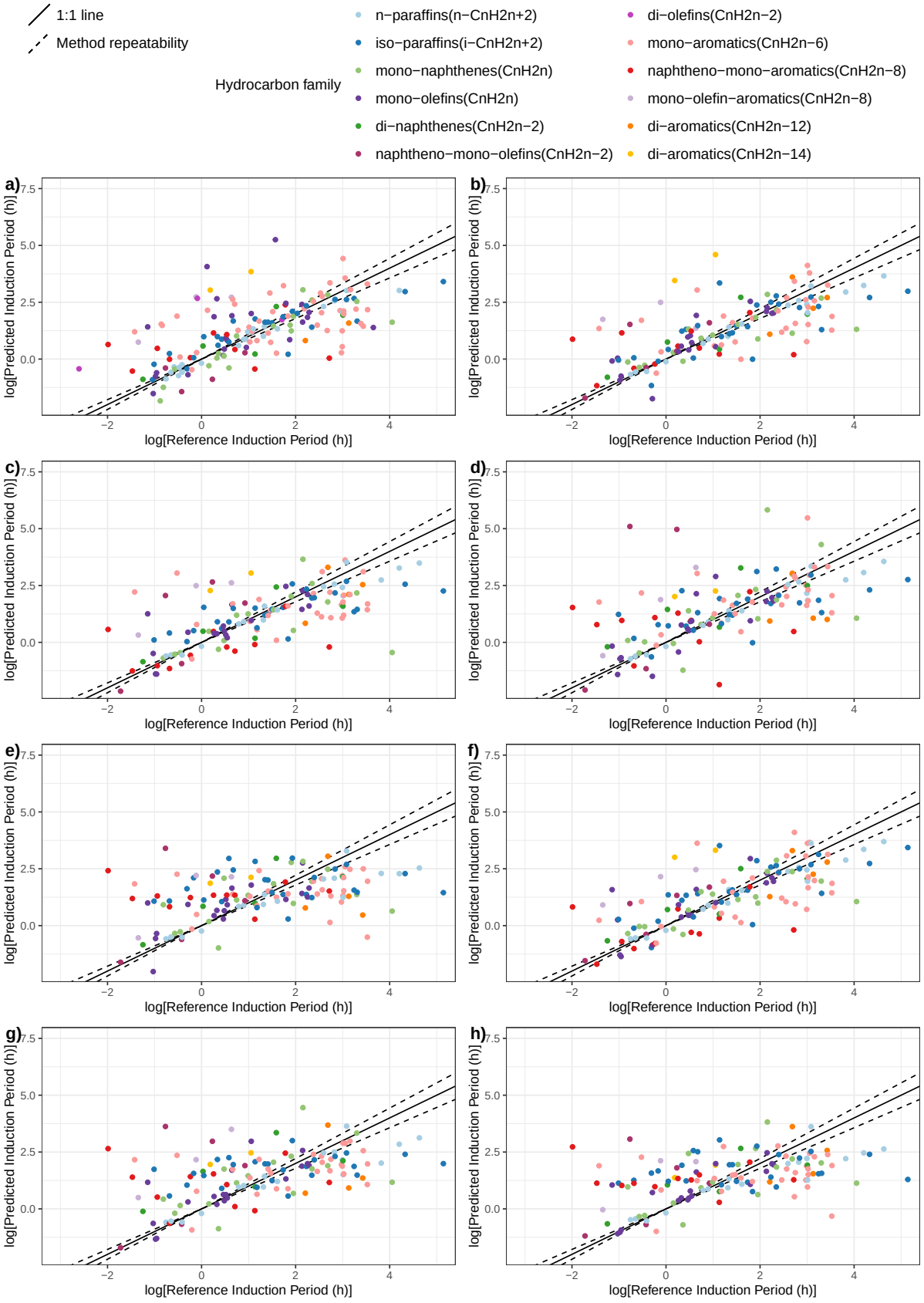


FIGURE 4.22 – Reference vs. predicted IP values in log scale for the 5 external CV folds. **a)** No pre-processing, **b)** Iteratively reweighted least squares (baseline correction), **c)** MinMax scaling, **d)** Savitzy-Golay 1st derivative (SG-1), **e)** Savitzy-Golay 2nd derivative (SG-2), **f)** Baseline correction + Minmax scaling, **g)** Baseline correction + MinMax scaling + SG1 **h)** Baseline correction + MinMax scaling + SG2

Near-Infrared spectroscopy has some important limitations when applied to the prediction of oxidation stability. As discussed in Chapter 2, bonding patterns play a critical role in determining the stability of aromatics, paraffins, and olefins. For example, compounds with tertiary benzylic carbons, such as cumene and *sec*-butylbenzene, are highly reactive, with induction periods of 0.2 h and 2.7 h, respectively. In contrast, *tert*-butylbenzene, which lacks a tertiary benzylic carbon, is significantly more stable, with an IP of 29.6 h. However, these very important molecular features cannot be observed in NIR spectroscopy. Quaternary carbons cannot be detected, while tertiary carbons (methine groups) exhibit weak absorption, which overlaps with more intense signals from methyl and methylene groups (see figure 4.23).

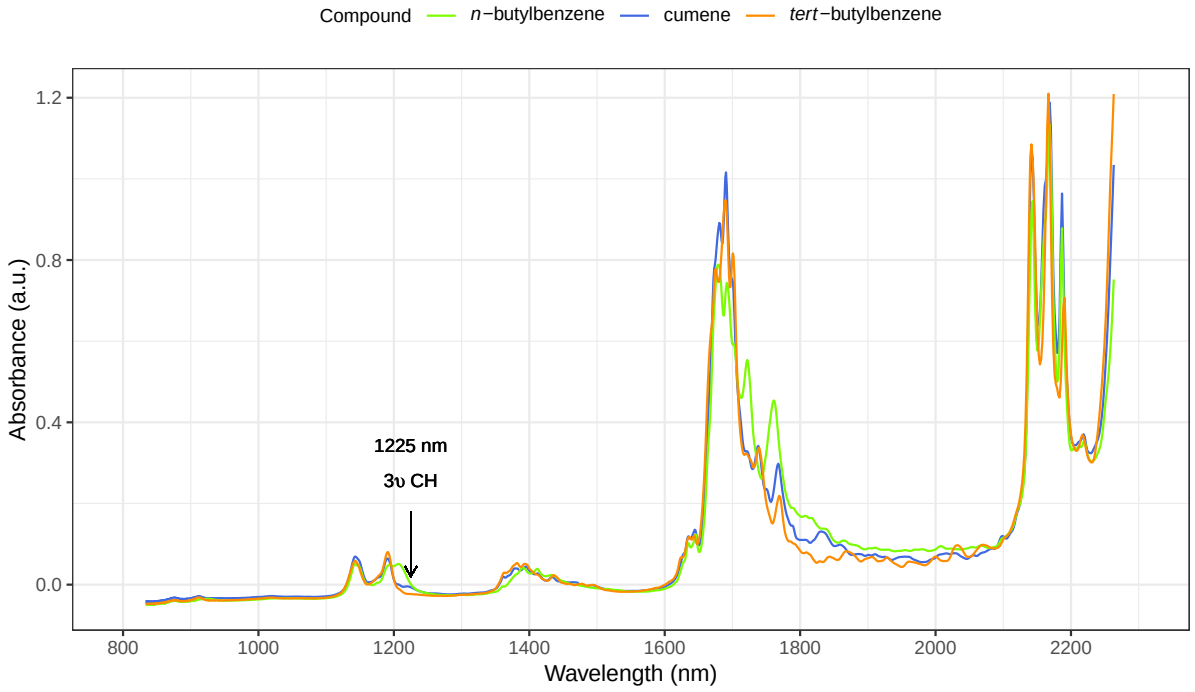


FIGURE 4.23 – NIR spectra for *n*-butylbenzene, cumene, and *tert*-butylbenzene, from 833 to 2265 nm. Cumene was selected instead of *sec*-butylbenzene for better visualization of the methine band.

4.5.3 NMR spectra-based models

Based on the *a posteriori* knowledge of the molecular features influencing oxidation stability, we propose that NMR spectroscopy could serve as a promising alternative for modeling the Induction Period. Unfortunately, due to time constraints, it was not possible to acquire experimental NMR spectra for this study. However, the NMRium tool [309, 310] provides the capability to simulate ^{1}H and ^{13}C NMR spectra. While simulated data may not fully replicate the real NMR data, it can still serve as a valuable "proof of concept" to explore the potential of this spectroscopic technique for predictive modeling.

In this work, we simulated the ^{1}H and ^{13}C NMR spectra for the 95 hydrocarbons in our database. For the simulations, we used the following parameters:

- Frequency: 1200 MHz
- ^{1}H range: -1 to 12 ppm
- ^{13}C range: -5 to 220 ppm
- Line width: 1 Hz
- Number of points: 128 000

In figure 4.24, we present the proton and carbon NMR spectra of two molecules poorly modeled by the NIR approach: cumene and *tert*-butylbenzene. The figure highlights the superior utility of NMR spectroscopy over NIR for distinguishing molecules with tertiary and quaternary carbons. For instance, the methine group exhibits a signal at approximately 3 ppm in the proton spectrum. While quaternary carbons lack a corresponding signal in this region, they are identifiable in the carbon spectrum with a signal at approximately 36 ppm. Thus, the presence of a quaternary carbon can be inferred from the absence of a proton signal coupled with the presence of its characteristic carbon signal.

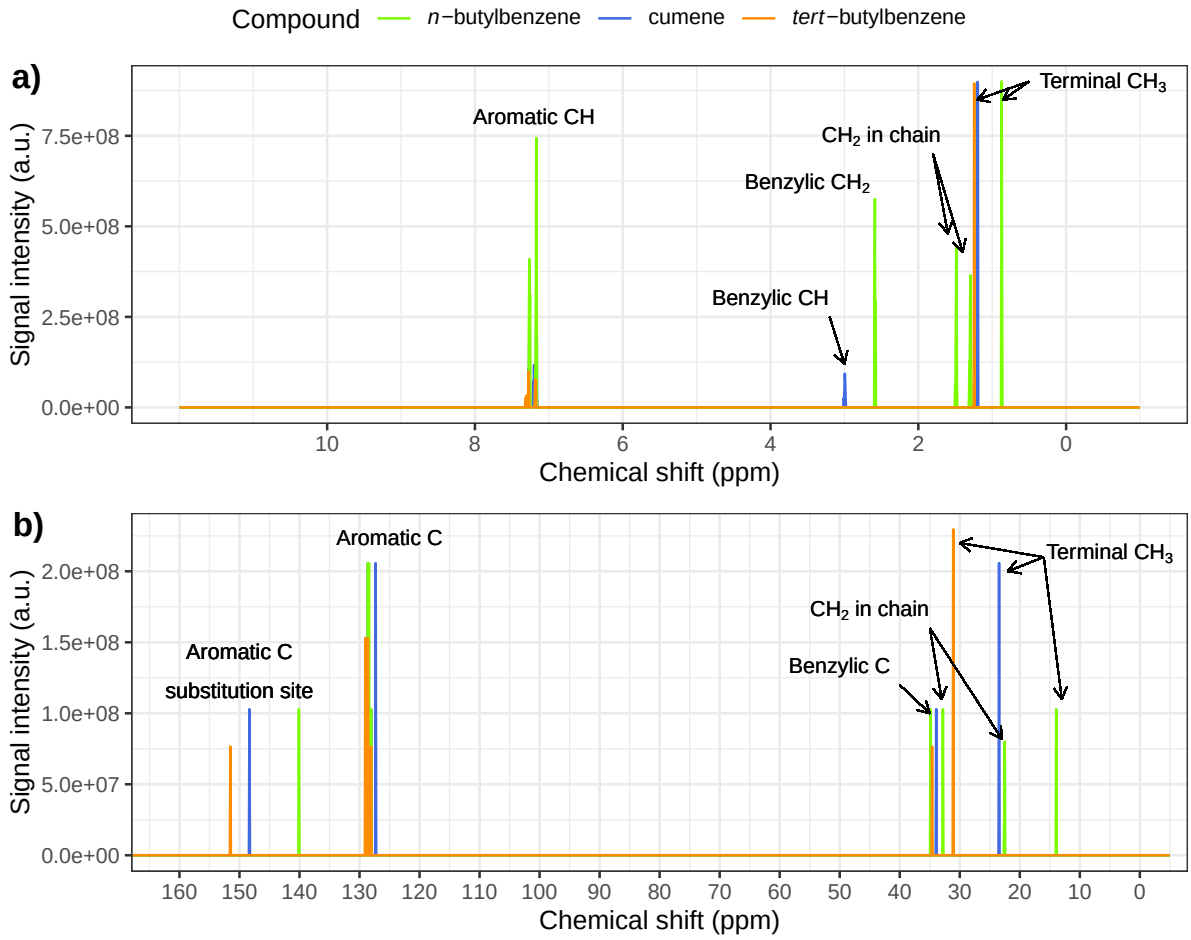


FIGURE 4.24 – **a)** ^{1}H and **b)** ^{13}C NMR spectra of *n*-butylbenzene, cumene, and *tert*-butylbenzene.

Interestingly, the chemical shift itself provides valuable insight into molecular features. Signals associated with paraffinic atoms show the lowest chemical shifts, followed by those of benzylic atoms, while aromatic atoms exhibit the highest chemical shifts. This distinction makes NMR a powerful tool for characterizing structural differences relevant to oxidation stability.

4.5.3.1 Data visualization and pre-processing

Before modeling, we decided to first visualize the data to see if there was relevant chemical information. For this, we first applied MinMax scaling to the proton and carbon NMR spectra, separately, as recommended by Leniak et al. [311]. We then concatenated the two data matrices, and proceeded to perform PCA.

As illustrated in figure 4.25, PC1 primarily distinguishes between naphthenes, aromatics, and linear paraffins, while PC2 discriminates molecules based on size, as evidenced by the vertical patterns formed by naphthenes, linear alkanes, and aromatics. The relatively low variance explained by these two PCs may stem from the excessively high resolution of the dataset, which comprises 131,080 columns.

A straightforward approach to address this dimensionality issue while preserving the overall data structure is the application of bucketing [311]. This pre-processing technique reduces the resolution by grouping adjacent variables and summing their signals, resulting in a simplified spectrum with fewer variables that retains the essential features of the original data. The scores plot for the bucketed data (bucket size = 500) is presented in figure 4.26.

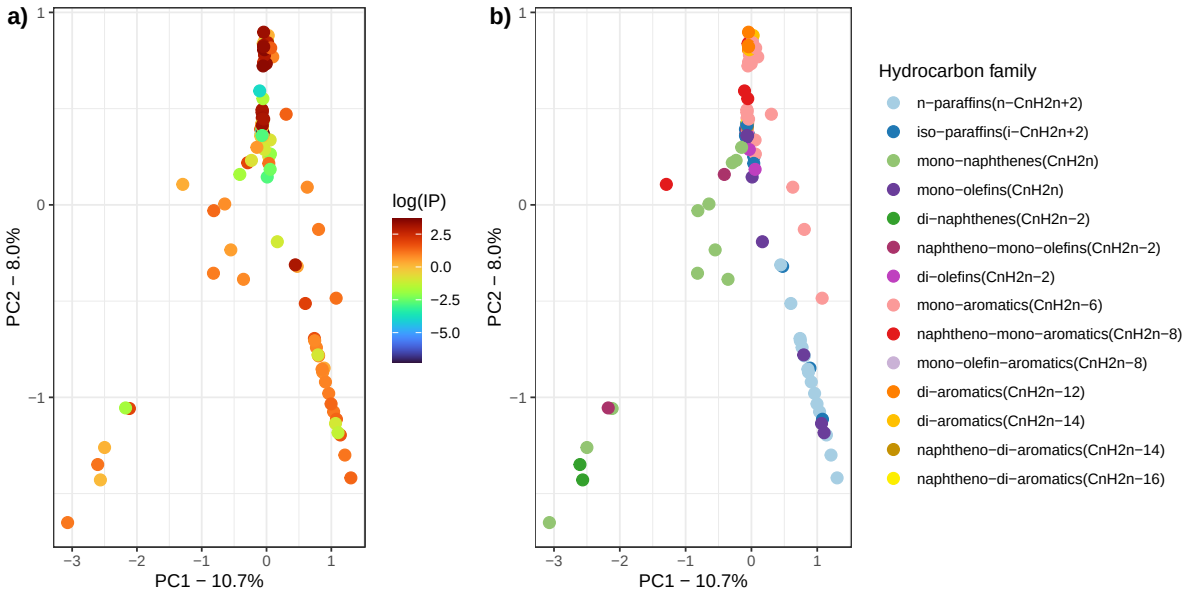


FIGURE 4.25 – Scores plot for PC1 vs. PC2 based on the simulated NMR spectra after scaling. The data points are color-coded to represent **a)** log(IP) values and **b)** hydrocarbon families.

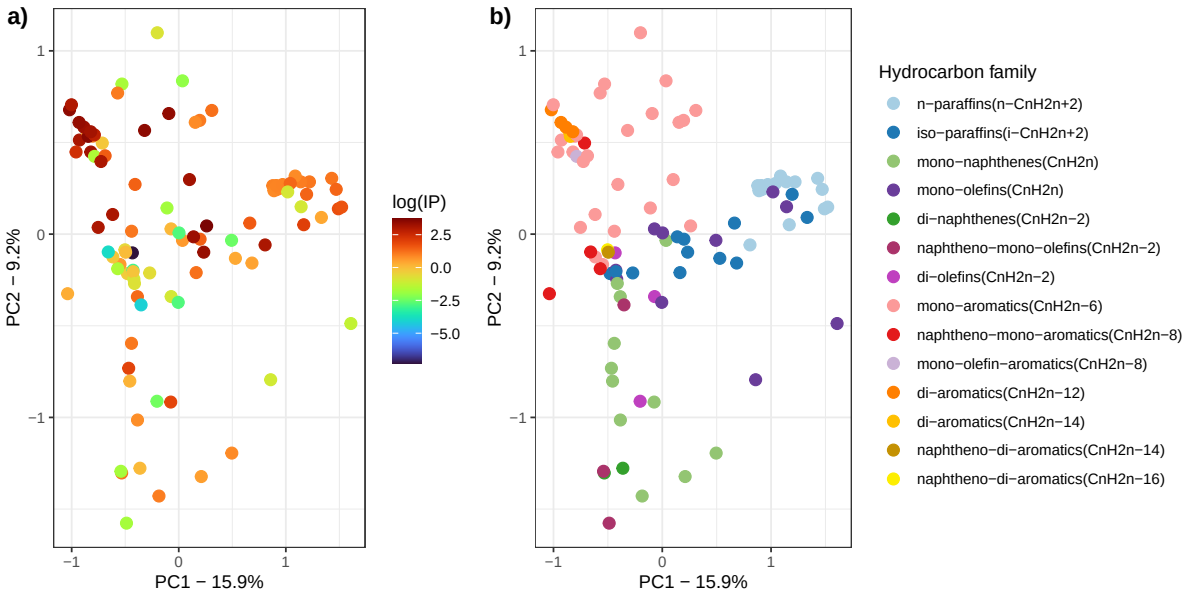


FIGURE 4.26 – Scores plot for PC1 vs. PC2 based on the simulated scaled NMR spectra after scaling and bucketing. The data points are color-coded to represent **a)** log(IP) values and **b)** hydrocarbon families.

4.5.3.2 Model performance

For model development, we employed the same methodology outlined in section 4.4.2, with one key difference: all spectra were scaled, and various bucket sizes were tested. A bucket size of 500 was found to be the most effective for optimizing model performance. Given the smaller dataset size compared to the NIR dataset, we were able to utilize three modeling algorithms: SVR-RBF, XGBoost Tree, and XGBoost Linear. The goodness-of-fit statistics for these models are summarized in table 4.3. The XGBoost Linear model achieves the lowest RMSE, with a relative error of 125%. In comparison, the SVR-RBF and XGBoost Tree models yield relative errors of 138% and 131%, respectively. The error of the XGBoost Linear model is comparable to that of the best QSPR-based model without data augmentation, which has a relative error of 123%. In contrast, the NIR-based model demonstrates a significantly higher relative error of 161%.

TABLE 4.3 – Model performance metrics for SVR-RBF, XGBoost with linear learners and XGBoost with regression trees learners. RMSE values were averaged across the 5 external cross-validation folds.

Model	**Optimal hyper-parameters**	**Scale**	**RMSEC**	**RMSECV**	**RMSEP**
SVR-RBF	$\sigma = 0.05$ $C = 100.00$ $\epsilon = 0.25$	log hour	0.13 0.91	1.26 5.38	1.81 10.41
XGBoost Linear	n_rounds = 80 eta = 0.001 $\lambda = 2.5$ $\alpha = 3.0$	log hour	0.60 3.63	1.03 5.81	1.20 7.04
XGBoost Tree	n_rounds = 150 eta = 0.1 max_depth = 5 $\gamma = 0$ colsample_bytree = 1 min_child_weight=0.50 subsample = 0.66	log hour	0.82 5.03	1.44 7.93	1.72 9.87

These results suggest that NMR spectra capture more molecular features related to oxidation stability than NIR spectra, which, as previously discussed, are "blind" to certain key molecular characteristics. However, as illustrated in figure 4.27, the XGBoost Linear model based on NMR data still struggles to accurately predict the behavior of certain aromatic and alkene compounds.

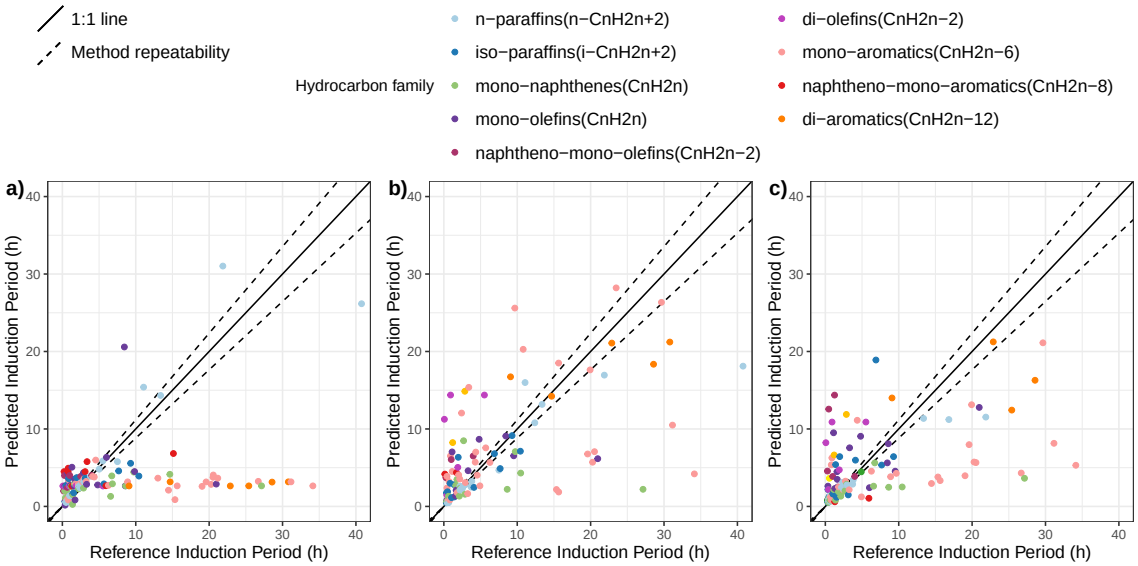


FIGURE 4.27 – Reference vs. predicted IP values for the 5 external CV folds of the NMR spectra-based model. Shown for **a)** Support Vector Regression, **b)** XGBoost with linear learners, and **c)** XGBoost with regression trees.

Future work could involve acquiring experimental NMR spectra to validate the results. Additionally, combining molecular descriptors with NMR spectra for model development could be an intriguing avenue for exploration. This hybrid approach might lead to the creation of a more accurate and robust predictive model.

4.6 Conclusions

In this chapter, we discussed the results from spectra-based modeling of the Induction Period. As highlighted earlier, the model derived from scaled NIR spectra using the SVR-RBF algorithm presented very low accuracy, with a relative error of 161%. This performance was inferior even to the least accurate non-data-augmented QSPR-based model.

We attribute this limitation primarily to the inherent constraints of NIR spectroscopy. Specifically, NIR cannot detect quaternary carbons, and methine groups exhibit minimal absorption, making it nearly impossible for the model to account for these critical features associated with oxidation stability. Additionally, the stability of tetra-substituted alkenes is significantly overestimated by the NIR-based model because their spectra do not present the characteristic C-H bands of vinyl groups.

We discussed that any spectroscopic technique used for modeling oxidation stability must be sensitive to the molecular features directly influencing this property, even when applying advanced machine-learning algorithms. For this reason, we suggest the use of ^{1}H and ^{13}C NMR spectroscopy. NMR provides vast molecular information, including the ability to directly identify methine groups and infer the presence of quaternary carbons through the combined analysis of proton and carbon spectra. Moreover, the chemical shift values at which signals are observed provide useful information, permitting to differentiate between atoms in paraffinic chains, benzylic sites, or aromatic rings.

Incorporating a spectroscopic technique like NMR could complement the information provided by QSPR-based models, potentially enhancing the overall accuracy of oxidation stability predictions.

Chapter 5

General conclusions and perspectives

As mentioned in section 1.5, this thesis had the following objectives:

- Generate a database from accelerated oxidation measurements of pure hydrocarbons.
- Perform spectroscopic measurements of fresh and oxidized hydrocarbons.
- Develop a cheminformatic model based on molecular descriptors to predict the oxidation stability.
- Develop a chemometric model based on the spectroscopic signals to predict the oxidation stability.
- Identify relevant features, chemical descriptors and spectroscopic regions, related to oxidation stability.

Now, we will discuss the how each of these objectives was reached. First, in this thesis, we performed the most extensive study of hydrocarbon oxidation stability to date. This comprehensive study is of significant value to the scientific community, as the findings can serve as a foundation for deriving new reactivity trends and developing predictive models. Among the noteworthy results, we highlight the impact of ring strain on the oxidation stability of naphthenes, the role of substitution at the vinylic site in alkenes, and the effect of double activation of benzylic sites in di-aromatics.

For the QSPR-based model, we designed a robust set of 42 molecular descriptors informed by the experimental reactivity trends identified in this work. The results indicate that non-linear algorithms, such as XGBoost, are particularly well-suited for modeling oxidation stability. The resulting model achieved semi-quantitative accuracy, demonstrating its utility for screening hydrocarbons in the development of new jet-fuel candidates.

In contrast, the NIR spectroscopy-based model demonstrated low predictive accuracy, rendering it unsuitable for practical use. A key limitation of NIR spectroscopy is its inability to

capture certain critical molecular features associated with stability, such as bonding patterns. For instance, methine groups (tertiary carbons) exhibit low absorption, while quaternary carbons are undetectable in NIR spectra. The low accuracy of the resulting models rendered their interpretation of low interest.

As an alternative to NIR spectroscopy, we proposed the use of NMR spectra. However, due to time constraints, it was not possible to acquire experimental NMR data, so we suggested the use of simulated spectra instead. Thus, we developed a model based on simulated ^{1}H and ^{13}C NMR spectra, which achieved an accuracy comparable to the QSPR model. This improvement is attributed to the greater structural detail provided by NMR spectra, which captures essential molecular characteristics relevant to oxidation stability.

Future research in this area could explore several directions. On the experimental side, expanding the dataset beyond the 95 hydrocarbons analyzed in this study would be valuable for identifying additional structure-property relationships and improving the accuracy of the developed models. Increasing the diversity of compounds in the database would also provide greater insight into the reactivity of underrepresented hydrocarbon families. Moreover, investigating the oxidation stability of mixtures could yield a deeper understanding of non-linear blending effects, such as those observed when mixing paraffins with naphthalenes or olefins. These studies could utilize a surrogate mixture approach to replicate the chemical composition of jet fuel, potentially paving the way for more accurate predictive models for real jet fuels

On the modeling side, several options can be explored. One potential direction is rethinking the prediction methodology for the IP. For example, the IP range could be discretized into categories such as low, medium, and high stability, enabling classification-based modeling. However, such discretization may not be a trivial choice. Another promising avenue involves combining multiple data sources to enhance model accuracy. For instance, integrating molecular descriptors used in QSPR-based modeling with spectral data from techniques like NMR could provide a more comprehensive view of hydrocarbon behavior. Additionally, incorporating theoretical descriptors, such as gas-phase kinetic constants or bond dissociation energies, could further refine model performance. These advancements would not only deepen our understanding of oxidation stability but also support the development of new jet fuel candidates.

References

[1] A. N. Sarkar. Evolving Green Aviation Transport System: A Hoilistic Approah to Sustainable Green Market Development. *American Journal of Climate Change*, 01(03): 164–180, 2012. ISSN 2167-9495. doi: 10.4236/ajcc.2012.13014.

[2] Air Transport Action Group. Facts & figures, 06/04/2024. URL https://www.atag.org/facts-figures.

[3] Aviation Benefits Beyond Borders. Global fact sheet: Scope of aviation, 2020. URL https://www.iata.org/en/iata-repository/pressroom/fact-sheets/fact-sheet-benefits-aviation-statistics/.

[4] International Civil Aviation Organization. Future of Aviation, 06/04/2024. URL https://www.icao.int/Meetings/FutureOfAviation/Pages/default.aspx.

[5] C. Zhang, X. Hui, Y. Lin, and C.-J. Sung. Recent development in studies of alternative jet fuel combustion: Progress, challenges, and opportunities. *Renewable and Sustainable Energy Reviews*, 54:120–138, 2016. ISSN 13640321. doi: 10.1016/j.rser.2015.09.056.

[6] B. Graver, X. S. Zheng, D. Rutherford, J. Mukhopadhaya, and E. Pronk. Vision 2050: Aligning aviation with the Paris Agreement. *The International Council on Clean Transportation*, 2022.

[7] International Energy Agency: IEA. Aviation: Tracking Aviation, 06/04/2024. URL https://www.iea.org/energy-system/transport/aviation.

[8] J. Holladay, Z. Abdullah, and J. Heyne. Sustainable Aviation Fuel: Review of Technical Pathways. pages 1–4, 2020. doi: 10.2172/1660415.

[9] ASTM International. *Specification for Aviation Turbine Fuel Containing Synthesized Hydrocarbons (ASTM D7566-24d)*. ASTM International, West Conshohocken, PA 19428-2959, U.S.A., 2024. doi: 10.1520/D7566-24D. URL https://www.astm.org/d7566-24d.html.

[10] ASTM International. *Specification for Aviation Turbine Fuels (ASTM D1655-24b)*. ASTM International, West Conshohocken, PA 19428-2959, U.S.A., 2024. doi: 10.1520/D1655-24B. URL https://www.astm.org/d1655-24b.html.

[11] P. Vozka, B. A. Modereger, A. C. Park, W. T. J. Zhang, R. W. Trice, H. I. Kenttämaa, and G. Kilaz. Jet fuel density via GC × GC-FID. *Fuel*, 235:1052–1060, 2019. ISSN 00162361. doi: 10.1016/j.fuel.2018.08.110.

[12] J. T. Edwards. *Reference Jet Fuels for Combustion Testing*. 55th AIAA Aerospace Sciences Meeting, American Institute of Aeronautics and Astronautics, Grapevine, Texas, U.S.A., 2017. doi: 10.2514/6.2017-0146.

[13] T. Jia, X. Zhang, Y. Liu, S. Gong, C. Deng, L. Pan, and J.-J. Zou. A comprehensive review of the thermal oxidation stability of jet fuels. *Chemical Engineering Science*, 229:116157, 2021. ISSN 00092509. doi: 10.1016/j.ces.2020.116157.

[14] Air bp. What is sustainable aviation fuel (SAF)?, 2022. URL https://www.bp.com/en/global/air-bp/news-and-views/views/what-is-sustainable-aviation-fuel-saf-and-why-is-it-important.html.

[15] P. Vozka, D. Vrtiška, P. Šimáček, and G. Kilaz. Impact of Alternative Fuel Blending Components on Fuel Composition and Properties in Blends with Jet A. *Energy & Fuels*, 33(4):3275–3289, 2019. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.9b00105.

[16] A. Ben Amara, S. Kaoubi, and L. Starck. Toward an optimal formulation of alternative jet fuels: Enhanced oxidation and thermal stability by the addition of cyclic molecules. *Fuel*, 173:98–105, 2016. ISSN 00162361. doi: 10.1016/j.fuel.2016.01.040.

[17] J. Pullen and K. Saeed. An overview of biodiesel oxidation stability. *Renewable and Sustainable Energy Reviews*, 16(8):5924–5950, 2012. ISSN 13640321. doi: 10.1016/j.rser.2012.06.024.

[18] Y. Chang, M. Jia, Y. Li, Y. Liu, M. Xie, H. Wang, and R. D. Reitz. Development of a skeletal mechanism for diesel surrogate fuel by using a decoupling methodology. *Combustion and Flame*, 162(10):3785–3802, 2015. ISSN 00102180. doi: 10.1016/j.combustflame.2015.07.016.

[19] W. Yu, W. Yang, K. Tay, and F. Zhao. Development of a new skeletal mechanism for decalin oxidation under engine relevant conditions. *Fuel*, 212:41–48, 2018. ISSN 00162361. doi: 10.1016/j.fuel.2017.10.009.

[20] W. Yu, F. Zhao, W. Yang, K. Tay, and H. Xu. Development of an optimization methodology for formulating both jet fuel and diesel fuel surrogates and their associated skeletal oxidation mechanisms. *Fuel*, 231:361–372, 2018. ISSN 00162361. doi: 10.1016/j.fuel.2018.05.121.

[21] M. H. Abraham. Application of solvation equations to chemical and biochemical processes. *Pure and Applied Chemistry*, 65(12):2503–2512, 1993. ISSN 0033-4545. doi: 10.1351/pac199365122503.

[22] C. Mintz, T. Ladlie, K. Burton, M. Clark, W. E. Acree, and M. H. Abraham. Enthalpy of Solvation Correlations for Gaseous Solutes Dissolved in Alcohol Solvents based on the Abraham Model. *QSAR & Combinatorial Science*, 27(5):627–635, 2008. ISSN 1611020X. doi: 10.1002/qsar.200730128.

[23] M. D. Le, V. Warth, L. Giarracca, E. Moine, R. Bounaceur, R. Privat, J.-N. Jaubert, R. Fournet, P.-A. Glaude, and B. Sirjean. Development of a Detailed Kinetic Model for the Oxidation of n-Butane in the Liquid Phase. *The journal of physical chemistry. B*, 125(25):6955–6967, 2021. doi: 10.1021/acs.jpcb.1c02988.

[24] B. H. H. Goh, C. T. Chong, H. C. Ong, T. Seljak, T. Katrašnik, V. Józsa, J.-H. Ng, B. Tian, S. Karmarkar, and V. Ashokkumar. Recent advancements in catalytic conversion pathways for synthetic jet fuel produced from bioresources. *Energy Conversion and Management*, 251:114974, 2022. ISSN 01968904. doi: 10.1016/j.enconman.2021.114974.

[25] M. R. Riazi. *Characterization and properties of petroleum fractions*. ASTM International, W. Conshohocken PA, 2005. ISBN 0803133618.

[26] M. Commodo, I. Fabris, C. P. T. Groth, and Ö. L. Gülder. Analysis of Aviation Fuel Thermal Oxidative Stability by Electrospray Ionization Mass Spectrometry (ESI–MS). *Energy & Fuels*, 25(5):2142–2150, 2011. ISSN 0887-0624. doi: 10.1021/ef2002102.

[27] R. C. Striebich, J. Contreras, L. M. Balster, Z. West, L. M. Shafer, and S. Zabarnick. Identification of Polar Species in Aviation Fuels using Multidimensional Gas Chromatography-Time of Flight Mass Spectrometry. *Energy & Fuels*, 23(11):5474–5482, 2009. ISSN 0887-0624. doi: 10.1021/ef900386x.

[28] G. W. Mushrush, E. J. Beal, J. M. Hughes, S. E. Bonde, W. L. Gore, and G. E. Dolbear. STABILITY STUDIES OF A JET FUEL CONTAINING NO ORGANO- SULFUR COMPOUNDS. *Petroleum Science and Technology*, 20(5-6):561–570, 2002. ISSN 1091-6466. doi: 10.1081/LFT-120003580.

[29] ExxonMobil. ExxonMobil Jet A-1. URL https://www.exxonmobil.com/en/aviation/products-and-services/products/exxonmobil-jet-a-1.

[30] National Aviation Academy. What are the different types of aviation fuel? URL https://www.naa.edu/aviation-fuel/.

[31] Shell Global. Civil jet fuel: Grades and specifications, . URL https://www.shell.com/business-customers/aviation/aviation-fuel/civil-jet-fuel-grades.html.

[32] Shell Global. Military jet fuel, . URL https://aviation.totalenergies.com/en/fuels-and-services-aviation/aviation-fuels/jet-a1.

[33] Repsol Global. Kerosene JP-8 - fuel for military aircraft. URL https://www.repsol.com/en/products-and-services/aviation/jp-8/index.cshtml.

[34] U.S. Department of Energy. Sustainable Aviation Fuel, 2022. URL https://afdc.energy.gov/fuels/sustainable-aviation-fuel.

[35] R. S. Capaz, J. A. Posada, P. Osseweijer, and J. E. Seabra. The carbon footprint of alternative jet fuels produced in Brazil: exploring different approaches. *Resources, Conservation and Recycling*, 166:105260, 2021. ISSN 09213449. doi: 10.1016/j.resconrec.2020.105260.

[36] P. Kurzawska. Overview of Sustainable Aviation Fuels including emission of particulate matter and harmful gaseous exhaust gas compounds. *Transportation Research Procedia*, 59:38–45, 2021. ISSN 23521465. doi: 10.1016/j.trpro.2021.11.095.

[37] A. L. Lown, L. Peereboom, S. A. Mueller, J. E. Anderson, D. J. Miller, and C. T. Lira. Cold flow properties for blends of biofuels with diesel and jet fuels. *Fuel*, 117:544–551, 2014. ISSN 00162361. doi: 10.1016/j.fuel.2013.09.067.

[38] M. E. Dry. Fischer–Tropsch reactions and the environment. *Applied Catalysis A: General*, 189(2):185–190, 1999. ISSN 0926860X. doi: 10.1016/S0926-860X(99)00275-6.

[39] L. Starck, L. Pidol, N. Jeuland, T. Chapus, P. Bogers, and J. Bauldreay. Production of Hydroprocessed Esters and Fatty Acids (HEFA) – Optimisation of Process Yield. *Oil & Gas Science and Technology – Revue d'IFP Energies nouvelles*, 71(1):10, 2016. ISSN 1294-4475. doi: 10.2516/ogst/2014007.

[40] N. Harich, R. Bassou, M. W. Priddy, T. E. Lacy, C. U. Pittman, and S. Kundu. Effects of alternative jet fuel blends on aerospace-grade carbon/epoxy composites. *Materials & Design*, 221:110993, 2022. ISSN 02641275. doi: 10.1016/j.matdes.2022.110993.

[41] S. Geleynse, K. Brandt, M. Garcia-Perez, M. Wolcott, and X. Zhang. The Alcohol-to-Jet Conversion Pathway for Drop-In Biofuels: Techno-Economic Evaluation. *ChemSusChem*, 11(21):3728–3741, 2018. doi: 10.1002/cssc.201801690.

[42] N. Nguyen and W. E. Tyner. Assessment of the feasibility of the production of alternative jet fuel and diesel using catalytic hydrothermolysis technology: a stochastic techno–economic analysis. *Biofuels, Bioproducts and Biorefining*, 16(1):91–104, 2022. ISSN 1932-104X. doi: 10.1002/bbb.2258.

[43] J. Yang, Z. Xin, Q. He, K. Corscadden, and H. Niu. An overview on performance characteristics of bio-jet fuels. *Fuel*, 237:916–936, 2019. ISSN 00162361. doi: 10.1016/j.fuel.2018.10.079.

[44] A. P. P. Pires, Y. Han, J. Kramlich, and M. Garcia-Perez. Chemical Composition and Fuel Properties of Alternative Jet Fuels. *BioResources*, 13(2):2632–2657, 2018. doi: 10.15376/biores.13.2.2632-2657.

[45] R. Natelson, M. Kurman, D. Miller, and N. Cernansky. Oxidation of Alternative Jet Fuels and their Surrogate Components. In *46th AIAA Aerospace Sciences Meeting and Exhibit*, Reston, Virigina, 01072008. American Institute of Aeronautics and Astronautics. ISBN 978-1-62410-128-1. doi: 10.2514/6.2008-970.

[46] International Civil Aviation Organization. Conversion processes, 06/04/2024. URL https://www.icao.int/environmental-protection/GFAAF/Pages/conversion-processes.aspx.

[47] J. Pullen and K. Saeed. Experimental study of the factors affecting the oxidation stability of biodiesel FAME fuels. *Fuel Processing Technology*, 125:223–235, 2014. ISSN 03783820. doi: 10.1016/j.fuproc.2014.03.032.

[48] Z. H. Sander, Z. J. West, J. S. Ervin, and S. Zabarnick. Experimental and Modeling Studies of Heat Transfer, Fluid Dynamics, and Autoxidation Chemistry in the Jet Fuel Thermal Oxidation Tester (JFTOT). *Energy & Fuels*, 29(11):7036–7047, 2015. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.5b01679.

[49] H. N. Stephens and F. L. Roduta. Oxidation in the Benzene Series by Gaseous Oxygen. V. The Oxidation of Tertiary Hydrocarbons. *Journal of the American Chemical Society*, 57 (12):2380–2381, 1935. ISSN 0002-7863. doi: 10.1021/ja01315a015.

[50] R. G. Larsen, R. E. Thorpe, and F. A. Armfield. Oxidation Characteristics of Pure Hydrocarbons. *Industrial & Engineering Chemistry*, 34(2):183–193, 1942. ISSN 0019-7866. doi: 10.1021/ie50386a012.

[51] J. Carstensen, P. Vanrolleghem, W. Rauch, and P. Reichert. Terminology and methodology in modelling for water quality management — a discussion starter. *Water Science and Technology*, 36(5):157–168, 1997. ISSN 0273-1223. doi: 10.1016/S0273-1223(97)00470-8.

[52] S. T. Glad. Modeling of Dynamic Systems from First Principles. In J. Baillieul and T. Samad, editors, *Encyclopedia of Systems and Control*, pages 1–9. Springer London, London, 2013. ISBN 978-1-4471-5102-9. doi: 10.1007/978-1-4471-5102-9{\textunderscore}102-1.

[53] M. K. Habib, S. A. Ayankoso, and F. Nagata. Data-Driven Modeling: Concept, Techniques, Challenges and a Case Study. In *2021 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 1000–1007. IEEE, 8/8/2021 - 8/11/2021. ISBN 978-1-6654-4101-8. doi: 10.1109/ICMA52036.2021.9512658.

[54] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99:101805, 2023. ISSN 15662535. doi: 10.1016/j.inffus.2023.101805.

[55] R. Aris. Prolegomena to the rational analysis of systems of chemical reactions. *Archive for Rational Mechanics and Analysis*, 19(2):81–99, 1965. ISSN 0003-9527. doi: 10.1007/BF00282276.

[56] A. N. Gorban and G. S. Yablonsky. Three Waves of Chemical Dynamics. *Mathematical Modelling of Natural Phenomena*, 10(5):1–5, 2015. ISSN 0973-5348. doi: 10.1051/mmnp/201510501.

[57] E. T. Denisov and I. B. Afanas'ev. *Oxidation and antioxidants in organic chemistry and biology*. Taylor & Francis, Boca Raton FL, 2005. ISBN 0824753569.

[58] M. Alves-Fortunato, A. Baroni, L. Neocel, M. Chardin, M. Matrat, C. Boucaud, and M. Mazarin. Gasoline Oxidation Stability: Deposit Formation Tendencies Evaluated by PetroOxy and Autoclave Methods and GDI/PFI Engine Tests. *Energy & Fuels*, 35(22):18430–18440, 2021. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.1c02466.

[59] R. Farhoosh, R. Niazmand, M. Rezaei, and M. Sarabi. Kinetic parameter determination of vegetable oil oxidation under Rancimat test conditions. *European Journal of Lipid Science and Technology*, 110(6):587–592, 2008. ISSN 1438-7697. doi: 10.1002/ejlt.200800004.

[60] H. Kwon, A. Lele, J. Zhu, C. S. McEnally, L. D. Pfefferle, Y. Xuan, and A. C. van Duin. ReaxFF-based molecular dynamics study of bio-derived polycyclic alkanes as potential alternative jet fuels. *Fuel*, 279:118548, 2020. ISSN 00162361. doi: 10.1016/j.fuel.2020.118548.

[61] W. W. Focke, I. van der Westhuizen, and X. Oosthuysen. Biodiesel oxidative stability from Rancimat data. *Thermochimica Acta*, 633:116–121, 2016. ISSN 00406031. doi: 10.1016/j.tca.2016.03.023.

[62] K. Bacha, A. Ben-Amara, A. Vannier, M. Alves-Fortunato, and M. Nardin. Oxidation Stability of Diesel/Biodiesel Fuels Measured by a PetroOxy Device and Characterization of Oxidation Products. *Energy & Fuels*, 29(7):4345–4355, 2015. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.5b00450.

[63] S. P. Heneghan and S. Zabarnick. Oxidation of jet fuels and the formation of deposit. *Fuel*, 73(1):35–43, 1994. ISSN 00162361. doi: 10.1016/0016-2361(94)90185-6.

[64] N. J. Kuprowicz, S. Zabarnick, Z. J. West, and J. S. Ervin. Use of Measured Species Class Concentrations with Chemical Kinetic Modeling for the Prediction of Autoxidation and Deposition of Jet Fuels. *Energy & Fuels*, 21(2):530–544, 2007. ISSN 0887-0624. doi: 10.1021/ef060391o.

[65] D. J. Carlsson and J. C. Robb. Liquid-phase oxidation of hydrocarbons. Part 4.—Indene and tetralin: occurrence and mechanism of the thermal initiation reaction with oxygen. *Trans. Faraday Soc.*, 62(0):3403–3415, 1966. ISSN 0014-7672. doi: 10.1039/TF9666203403.

[66] F. Garcia-Ochoa, A. Romero, and J. Querol. Modeling of the thermal n-octane oxidation in the liquid phase. *Industrial & Engineering Chemistry Research*, 28(1):43–48, 1989. ISSN 0888-5885. doi: 10.1021/ie00085a009.

[67] S. Blaine and P. E. Savage. Reaction pathways in lubricant degradation. 3. Reaction model for n-hexadecane autoxidation. *Industrial & Engineering Chemistry Research*, 31 (1):69–75, 1992. ISSN 0888-5885. doi: 10.1021/ie00001a010.

[68] I. Hermans, J. Peeters, L. Vereecken, and P. A. Jacobs. Mechanism of thermal toluene autoxidation. *Chemphyschem : a European journal of chemical physics and physical chemistry*, 8(18):2678–2688, 2007. doi: 10.1002/cphc.200700563.

[69] J. Hoorn, J. van Soolingen, and G. F. Versteeg. Modelling toluene oxidation: incorporation of mass transfer phenomena. *Chemical Engineering Research and Design*, 83(2A):187–195, 2005. ISSN 0263-8762.

[70] S. Zabarnick. Chemical kinetic modeling of jet fuel autoxidation and antioxidant chemistry. *Industrial & Engineering Chemistry Research*, 32(6):1012–1017, 1993. ISSN 0888-5885. doi: 10.1021/ie00018a003.

[71] S. P. Heneghan and L. P. Chin. Autoxidation of jet fuels: Implications for modeling and thermal stability. *Proceedings of the 5th international conference on stability and handling of liquid fuels*, 1995. URL https://www.osti.gov/biblio/45051,journal=.

[72] E. Blurock and F. Battin-Leclerc. Modeling Combustion with Detailed Kinetic Mechanisms. In F. Battin-Leclerc, J. M. Simmie, and E. Blurock, editors, *Cleaner combustion*, Green Energy and Technology, pages 17–57. Springer, London, 2013. ISBN 978-1-4471-5306-1. doi: 10.1007/978-1-4471-5307-8{\textunderscore}2.

[73] Y. Chang, M. Jia, Y. Liu, Y. Li, M. Xie, and H. Yin. Application of a Decoupling Methodology for Development of Skeletal Oxidation Mechanisms for Heavy n -Alkanes from n -Octane to n -Hexadecane. *Energy & Fuels*, 27(6):3467–3479, 2013. ISSN 0887-0624. doi: 10.1021/ef400460d.

[74] P. Dagaut, A. Ristori, A. Frassoldati, T. Faravelli, G. Dayma, and E. Ranzi. Experimental and semi-detailed kinetic modeling study of decalin oxidation and pyrolysis over a wide range of conditions. *Proceedings of the Combustion Institute*, 34(1):289–296, 2013. ISSN 15407489. doi: 10.1016/j.proci.2012.05.099.

[75] W. Fan, M. Jia, Y. Chang, and M. Xie. Understanding the Relationship between Cetane Number and the Ignition Delay in Shock Tubes for Different Fuels Based on a Skeletal Primary Reference Fuel (n -Hexadecane/Iso-cetane) Mechanism. *Energy & Fuels*, 29(5): 3413–3427, 2015. ISSN 0887-0624. doi: 10.1021/ef5028185.

[76] D. C. Mielczarek, M. Matrat, A. B. Amara, Y. Bouyou, P. Wund, and L. Starck. Toward the Accurate Prediction of Liquid Phase Oxidation of Aromatics: A Detailed Kinetic Mechanism for Toluene Autoxidation. *Energy & Fuels*, 31(11):12893–12913, 2017. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.7b00416.

[77] S. Dong, S. W. Wagnon, L. P. Maffei, G. Kukkadapu, A. Nobili, Q. Mao, M. Pelucchi, L. Cai, K. Zhang, M. Raju, T. Chatterjee, W. J. Pitz, T. Faravelli, H. Pitsch, P. K. Senecal, and H. J. Curran. A new detailed kinetic model for surrogate fuels: C3MechV3.3. *Applications in Energy and Combustion Science*, 9:100043, 2022. ISSN 2666-352X. doi: 10.1016/j.jaecs.2021.100043. URL https://www.sciencedirect.com/science/article/pii/S2666352X21000212.

[78] L. J. Broadbelt, S. M. Stark, and M. T. Klein. Computer Generated Pyrolysis Modeling: On-the-Fly Generation of Species, Reactions, and Rates. *Industrial & Engineering Chemistry Research*, 33(4):790–799, 1994. ISSN 0888-5885. doi: 10.1021/ie00028a003.

[79] E. Ranzi, T. Faravelli, P. Gaffuri, and A. Sogaro. Low-temperature combustion: Automatic generation of primary oxidation reactions and lumping procedures. *Combustion and Flame*, 102(1-2):179–192, 1995. ISSN 00102180. doi: 10.1016/0010-2180(94)00253-O.

[80] V. Warth, N. Stef, P. A. Glaude, F. Battin-Leclerc, G. Scacchi, and G. M. Côme. Computer-Aided Derivation of Gas-Phase Oxidation Mechanisms: Application to the Modeling of the Oxidation of n-Butane. *Combustion and Flame*, 114(1-2):81–102, 1998. ISSN 00102180. doi: 10.1016/S0010-2180(97)00273-3.

[81] N. M. Vandewiele, K. M. van Geem, M.-F. Reyniers, and G. B. Marin. Genesys: Kinetic model construction using chemo-informatics. *Chemical Engineering Journal*, 207-208: 526–538, 2012. ISSN 1385-8947. doi: 10.1016/j.cej.2012.07.014. URL https://www.sciencedirect.com/science/article/pii/S1385894712009059.

[82] C. W. Gao, J. W. Allen, W. H. Green, and R. H. West. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Computer Physics Communications*, 203:212–225, 2016. ISSN 00104655. doi: 10.1016/j.cpc.2016.02.013.

[83] S. Miertuš, E. Scrocco, and J. Tomasi. Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chemical Physics*, 55(1):117–129, 1981. ISSN 0301-0104. doi: 10.1016/0301-0104(81)85090-2. URL https://www.sciencedirect.com/science/article/pii/0301010481850902.

[84] A. Klamt and G. Schüürmann. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2*, (5):799–805, 1993. ISSN 0300-9580. doi: 10.1039/P29930000799.

[85] A. Jalan, R. H. West, and W. H. Green. An extensible framework for capturing solvent effects in computer generated kinetic models. *The journal of physical chemistry. B*, 117 (10):2955–2970, 2013. doi: 10.1021/jp310824h.

[86] E. Voutsas, V. Louli, C. Boukouvalas, K. Magoulas, and D. Tassios. Thermodynamic property calculations with the universal mixing rule for EoS/GE models: Results with the Peng–Robinson EoS and a UNIFAC model. *Fluid Phase Equilibria*, 241(1-2):216–228, 2006. ISSN 03783812. doi: 10.1016/j.fluid.2005.12.028.

[87] ASTM International. *Test Method for Oxidation Stability of Middle Distillate Fuels-Rapid Small Scale Oxidation Test (RSSOT) (ASTM D7545-14(2019)e1)*. ASTM International, West Conshohocken, PA 19428-2959, U.S.A., 2019. doi: 10.1520/D7545-14R19E01. URL https://www.astm.org/d7545-14r19e01.html.

[88] ASTM International. *Test Method for Oxidation Stability of Spark Ignition Fuel-Rapid Small Scale Oxidation Test (RSSOT) (ASTM D7525-14(2019)e1)*. ASTM International, West Conshohocken, PA 19428-2959, U.S.A., 2019. doi: 10.1520/D7525-14R19E01. URL https://www.astm.org/d7525-14r19e01.html.

[89] M. Skolniak, P. Bukrejewski, and J. Frydrych. Analysis of Changes in the Properties of Selected Chemical Compounds and Motor Fuels Taking Place During Oxidation Processes. In K. Biernat, editor, *Storage Stability of Fuels*, pages 205–240. InTech, 2015. ISBN 978-953-51-1734-6. doi: 10.5772/59805.

[90] K. Chatelain, A. Nicolle, A. Ben Amara, L. Catoire, and L. Starck. Wide Range Experimental and Kinetic Modeling Study of Chain Length Impact on n -Alkanes Autoxidation. *Energy & Fuels*, 30(2):1294–1303, 2016. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.5b02470.

[91] K. Chatelain, A. Nicolle, A. Ben Amara, L. Starck, and L. Catoire. Structure–Reactivity Relationships in Fuel Stability: Experimental and Kinetic Modeling Study of Isoparaffin Autoxidation. *Energy & Fuels*, 32(9):9415–9426, 2018. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.8b01379.

[92] H. N. Stephens. Oxidations In The Benzene Series By Gaseous Oxygen I. Oxidation Of Methylbenzenes. *Journal of the American Chemical Society*, 48(7):1824–1826, 1926. ISSN 0002-7863. doi: 10.1021/ja01418a005.

[93] T. Edwards, editor. *Prospects for JP-8+225, a stepping stone to JP-900*. doi: 10.2514/6.1998-3532.

[94] J. Czarnocka, A. Matuszewska, and M. Odziemkowska. Autoxidation of Fuels During Storage. In *Storage Stability of Fuels*, pages 157–188. InTech, 2015. ISBN 9789535117346. doi: 10.5772/59807.

[95] J.-F. Poon and D. A. Pratt. Recent Insights on Hydrogen Atom Transfer in the Inhibition of Hydrocarbon Autoxidation. *Accounts of chemical research*, 51(9):1996–2005, 2018. doi: 10.1021/acs.accounts.8b00251.

[96] E. Alborzi, P. Gadsby, M. S. Ismail, A. Sheikhansari, M. R. Dwyer, A. J. H. M. Meijer, S. G. Blakey, and M. Pourkashanian. Comparative Study of the Effect of Fuel Deoxygenation and Polar Species Removal on Jet Fuel Surface Deposition. *Energy & Fuels*, 33(3):1825–1836, 2019. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.8b03468.

[97] T. Hartikka, U. Kiiski, M. Kuronen, and S. Mikkonen. Diesel Fuel Oxidation Stability: A Comparative Study. In *SAE Technical Paper Series*, SAE Technical Paper Series. SAE International400 Commonwealth Drive, Warrendale, PA, United States, 2013. doi: 10.4271/2013-01-2678.

[98] S. Gong, T. Jia, L. Pan, G. Nie, X. Zhang, Li Wang, and J.-J. Zou. Enhanced Thermal Oxidation Stability of Jet Fuel by Deoxygenation Treatment. *Chemistry and Technology of Fuels and Oils*, 56(4):627–637, 2020. ISSN 0009-3092. doi: 10.1007/s10553-020-01176-w.

[99] Z. J. West. *Studies of Jet Fuel Autoxidation Chemistry: Catalytic Hydroperoxide Decomposition & High Heat Flux Effects*. PhD thesis, University of Dayton, 2011. URL https://etd.ohiolink.edu/apexprod/rws_olink/r/1501/10?clear=10&p10_accession_num=dayton1322764905.

[100] L. Zhao, X. Zhang, L. Pan, and J. Liu. Storage period prediction and metal compatibility of endothermic hydrocarbon fuels. *Fuel*, 233:1–9, 2018. ISSN 00162361. doi: 10.1016/j.fuel.2018.06.034.

[101] A. Sarin, R. Arora, N. P. Singh, M. Sharma, and R. K. Malhotra. Influence of metal contaminants on oxidation stability of Jatropha biodiesel. *Energy*, 34(9):1271–1275, 2009. ISSN 03605442. doi: 10.1016/j.energy.2009.05.018.

[102] J. A. Waynick. The Development and Use of Metal Deactivators in the Petroleum Industry: A Review. *Energy & Fuels*, 15(6):1325–1340, 2001. ISSN 0887-0624. doi: 10.1021/ef010113j.

[103] I. Hermans, T. L. Nguyen, P. A. Jacobs, and J. Peeters. Autoxidation of Cyclohexane: Conventional Views Challenged by Theory and Experiment. *Chemphyschem : a European journal of chemical physics and physical chemistry*, 6(4):637–645, 2005. doi: 10.1002/cphc.200700563.

[104] C. G. Kabana, S. Botha, C. Schmucker, C. Woolard, and B. Beaver. Oxidative Stability of Middle Distillate Fuels. Part 1: Exploring the Soluble Macromolecular Oxidatively Reactive Species (SMORS) Mechanism with Jet Fuels. *Energy & Fuels*, 25(11):5145–5157, 2011. ISSN 0887-0624. doi: 10.1021/ef200964z.

[105] T. Jia, M. Zhao, L. Pan, C. Deng, J.-J. Zou, and X. Zhang. Effect of phenolic antioxidants on the thermal oxidation stability of high-energy–density fuel. *Chemical Engineering Science*, 247:117056, 2022. ISSN 00092509. doi: 10.1016/j.ces.2021.117056.

[106] W. A. Yehye, N. A. Rahman, A. Ariffin, S. B. Abd Hamid, A. A. Alhadi, F. A. Kadir, and M. Yaeghoobi. Understanding the chemistry behind the antioxidant activities of butylated hydroxytoluene (BHT): a review. *European journal of medicinal chemistry*, 101:295–312, 2015. doi: 10.1016/j.ejmech.2015.06.026.

[107] S. Aminane, M. Sicard, Y. Melliti, F. Ser, and L. Sicard. Experimental study of the kinetics of degradation of n-dodecane under thermo-oxidative stress at low temperature and mechanism inferred. *Fuel*, 307:121669, 2022. ISSN 00162361. doi: 10.1016/j.fuel.2021.121669.

[108] D. C. Mielczarek. *Autoxidation Behaviour of Hydrocarbons in the Context of Conventional and Alternative Aviation Fuels*. PhD thesis, 2015.

[109] H. H. Zuidema. Oxidation of lubricating oils. *Chemical reviews*, 38:197–226, 1946. doi: 10.1021/cr60120a001.

[110] J. Igarashi, R. K. Jensen, J. Lusztyk, S. Korcek, and K. U. Ingold. Autoxidation of alkylnaphthalenes. 2. Inhibition of the autoxidation of n-hexadecane at 160.degree.C. *Journal of the American Chemical Society*, 114(20):7727–7736, 1992. ISSN 0002-7863. doi: 10.1021/ja00046a019.

[111] T. von Kuegelgen. Characterization of Biodiesel Oxidation and Oxidation Products.

[112] J. P. Cosgrove, D. F. Church, and W. A. Pryor. The kinetics of the autoxidation of polyunsaturated fatty acids. *Lipids*, 22(5):299–304, 1987. ISSN 0024-4201. doi: 10.1007/BF02533996.

[113] G. Knothe and R. O. Dunn. Dependence of oil stability index of fatty compounds on their structure and concentration and presence of metals. *Journal of the American Oil Chemists' Society*, 80(10):1021–1026, 2003. ISSN 0003021X. doi: 10.1007/s11746-003-0814-x.

[114] ASTM International. *Test Method for Hydroperoxide Number of Aviation Turbine Fuels, Gasoline and Diesel Fuels (ASTM D3703-18)*. ASTM International, West Conshohocken, PA 19428-2959, U.S.A., 2024. doi: 10.1520/D3703-18. URL https://www.astm.org/d3703-18.html.

[115] H. Roohi and M. Rajabi. Iodometric Determination of Hydroperoxides in Hydrocarbon Autoxidation Reactions Using Triphenylphosphine Solution as a Titrant: A New Protocol. *Industrial & Engineering Chemistry Research*, 57(20):6805–6814, 2018. ISSN 0888-5885. doi: 10.1021/acs.iecr.7b05403.

[116] R. Benrabah, Z. El Sayah, M. Duy Le, Y. A. Derrick Warren, P.-A. Glaude, R. Fournet, and B. Sirjean. Experimental study of the impact of alcohols on the oxidation stability of a surrogate jet-fuel. *Fuel*, 361:130750, 2024. ISSN 00162361. doi: 10.1016/j.fuel.2023.130750. URL https://www.sciencedirect.com/science/article/pii/S0016236123033641.

[117] H. Shen, Y. Wang, J. Deng, L. Zhang, and Y. She. Catalyst-free and solvent-free oxidation of cycloalkanes (C5-C8) with molecular oxygen: Determination of autoxidation temperature and product distribution. *Chinese Journal of Chemical Engineering*, 26(5):1064–1070, 2018. ISSN 1004-9541. doi: 10.1016/j.cjche.2018.02.019. URL https://www.sciencedirect.com/science/article/pii/S1004954117314544.

[118] Y. Huang, F. Li, G. Bao, M. Li, and H. Wang. Qualitative and quantitative analysis of the influence of biodiesel fatty acid methyl esters on iodine value. *Environmental science and pollution research international*, 29(2):2432–2447, 2022. doi: 10.1007/s11356-021-15762-w.

[119] J. Cerny, Z. Strnad, G. Sebor. Composition and oxidation stability of SAE 15W-40 engine oils. *Tribology International*, 34, 2001. ISSN 0301679X.

[120] K. Engeländer, A. Duchowny, B. Blümich, and A. Adams. Analysis of Aging Products from Biofuels in Long-Term Storage. *ACS Omega*, 7(30):26256–26264, 2022. ISSN 2470-1343. doi: 10.1021/acsomega.2c01970.

[121] H. N. Stephens. Oxidations In The Benzene Series By Gaseous Oxygen II. Alkyl Benzenes With Two Or More Carbon Atoms In The Side Chain. *Journal of the American Chemical Society*, 48(11):2920–2922, 1926. ISSN 0002-7863. doi: 10.1021/ja01690a025.

[122] H. N. Stephens. Oxidation In The Benzene Series By Gaseous Oxygen III. Oxidation Of Alpha Phenyl Carbinols. *Journal of the American Chemical Society*, 50(1):186–190, 1928. ISSN 0002-7863. doi: 10.1021/ja01388a026.

[123] H. N. Stephens. Oxidation In The Benzene Series By Gaseous Oxygen. IV. Mechanism Of The Slow Oxidation Of Saturated Hydrocarbons. *Journal of the American Chemical Society*, 50(9):2523–2529, 1928. ISSN 0002-7863. doi: 10.1021/ja01396a031.

[124] S. Zabarnick. Studies of Jet Fuel Thermal Stability and Oxidation Using a Quartz Crystal Microbalance and Pressure Measurements. *Industrial & Engineering Chemistry Research*, 33(5):1348–1354, 1994. ISSN 0888-5885. doi: 10.1021/ie00029a034.

[125] S. Zabarnick and S. D. Whitacre. Aspects of Jet Fuel Oxidation. *Journal of Engineering for Gas Turbines and Power*, 120(3):519–525, 1998. ISSN 0742-4795. doi: 10.1115/1.2818177.

[126] S. Zabarnick and M. S. Mick. Inhibition of Jet Fuel Oxidation by Addition of Hydroperoxide-Decomposing Species. *Industrial & Engineering Chemistry Research*, 38(9):3557–3563, 1999. ISSN 0888-5885. doi: 10.1021/ie990107z.

[127] R. Benrabah, E. Girot, P. Arnoux, J.-M. Commenge, R. Fournet, P.-A. Glaude, and B. Sirjean. An innovative microfluidic reactor for testing the oxidation stability of fuels: application to a saf surrogate and comparison with the petrooxy test. In *IASH 2024, the 18TH INTERNATIONAL CONFERENCE ON STABILITY, HANDLING AND USE OF LIQUID FUELS*, Louisville (Kentucky), United States, 2024. URL https://hal.science/hal-04790643.

[128] F. Lacoste and L. Lagardere. Quality parameters evolution during biodiesel oxidation using Rancimat test. *European Journal of Lipid Science and Technology*, 105(3-4):149–155, 2003. ISSN 1438-7697. doi: 10.1002/ejlt.200390030.

[129] F. Bär, H. Hopf, M. Knorr, and J. Krahl. Rancimat and PetroOxy oxidation stability measurements of rapeseed oil methyl ester stabilized with hydrazides and antioxidants. *Fuel*, 232:108–113, 2018. ISSN 00162361. doi: 10.1016/j.fuel.2018.05.095.

[130] R. Mateos, M. Uceda, M. P. Aguilera, M. E. Escuderos, and G. Beltrán Maza. Relationship of Rancimat method values at varying temperatures for virgin olive oils. *European Food Research and Technology*, 223(2):246–252, 2006. ISSN 1438-2377. doi: 10.1007/s00217-005-0185-9.

[131] L. S. de Sousa, M. A. S. Garcia, E. C. P. Santos, J. do Nascimento Silva, A. G. de Castro, C. V. R. de Moura, and E. M. de Moura. Study of the kinetic and thermodynamic parameters of the oxidative degradation process of biodiesel by the action of antioxidants using the Rancimat and PetroOXY methods. *Fuel*, 238:198–207, 2019. ISSN 00162361. doi: 10.1016/j.fuel.2018.10.082.

[132] T. Prakoso, A. Tanaka, T. Hirotsu, P. Udomsap, N. Chollacoop, S. Goto, and A. Indarto. Oxidation stability of biodiesel fuel produced from Jatropha Curcas L using Rancimat and PetroOXY method. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 41(4):501–506, 2019. ISSN 1556-7036. doi: 10.1080/15567036.2018.1520333.

[133] X. Li, Y. Li, F. Yang, R. Liu, C. Zhao, Q. Jin, and X. Wang. Oxidation degree of soybean oil at induction time point under Rancimat test condition: Theoretical derivation and experimental observation. *Food research international (Ottawa, Ont.)*, 120:756–762, 2019. doi: 10.1016/j.foodres.2018.11.036.

[134] G. Datschefski. Hole of the JFTOT in aviation fuel stability research. *Fuel Science and Technology International*, 6(6):609–631, 1988. ISSN 0884-3759. doi: 10.1080/08843758808915906.

[135] S. G. Pande and D. R. Hardy. Effect of Copper, MDA, and Accelerated Aging on Jet Fuel Thermal Stability As Measured by the Gravimetric JFTOT. *Energy & Fuels*, 9(1):177–182, 1995. ISSN 0887-0624. doi: 10.1021/ef00049a026.

[136] J. Barker, J. Reid, S. Angel Smith, C. Snape, D. Scurr, G. Langley, K. Patel, A. Carter, C. Lapthorn, and F. Pullen. The Application of New Approaches to the Analysis of Deposits from the Jet Fuel Thermal Oxidation Tester (JFTOT). *SAE International Journal of Fuels and Lubricants*, 10(3), 2017. ISSN 1946-3960. doi: 10.4271/2017-01-2293.

[137] J. Zhou, Y. Xiong, and S. Xu. Evaluation of the oxidation stability of biodiesel stabilized with antioxidants using the PetroOXY method. *Fuel*, 184:808–814, 2016. ISSN 00162361. doi: 10.1016/j.fuel.2016.07.080.

[138] S. V. Araújo, B. S. Rocha, F. M. T. Luna, E. M. Rola, D. C. Azevedo, and C. L. Cavalcante. FTIR assessment of the oxidation process of castor oil FAME submitted to PetroOXY and Rancimat methods. *Fuel Processing Technology*, 92(5):1152–1155, 2011. ISSN 03783820. doi: 10.1016/j.fuproc.2010.12.026.

[139] M. L. Murta Valle, R. S. Leonardo, and J. Dweck. Comparative study of biodiesel oxidation stability using Rancimat, PetroOXY, and low P-DSC. *Journal of Thermal Analysis and Calorimetry*, 116(1):113–118, 2014. ISSN 1388-6150. doi: 10.1007/s10973-014-3706-6.

[140] F. Bär, M. Knorr, O. Schröder, H. Hopf, T. Garbe, and J. Krahl. Rancimat vs. rapid small scale oxidation test (RSSOT) correlation analysis, based on a comprehensive study of literature. *Fuel*, 291:120160, 2021. ISSN 00162361. doi: 10.1016/j.fuel.2021.120160.

[141] J. Beens. The role of gas chromatography in compositional analyses in the petroleum industry. *TrAC Trends in Analytical Chemistry*, 19(4):260–275, 2000. ISSN 01659936. doi: 10.1016/S0165-9936(99)00205-8.

[142] T. Potter and K. Simmons. *Composition of Petroleum Mixtures*. Total Petroleum Hydrocarbon Criteria Working Group Series. Amherst Scientific Publishing, 1998.

[143] K. Lissitsyna, S. Huertas, L. C. Quintero, and L. M. Polo. PIONA analysis of kerosene by comprehensive two-dimensional gas chromatography coupled to time of flight mass spectrometry. *Fuel*, 116:716–722, 2014. ISSN 00162361. doi: 10.1016/j.fuel.2013.07.077.

[144] A. Ben Amara, A. Nicolle, M. Alves-Fortunato, and N. Jeuland. Toward Predictive Modeling of Petroleum and Biobased Fuel Stability: Kinetics of Methyl Oleate/ n -Dodecane Autoxidation. *Energy & Fuels*, 27(10):6125–6133, 2013. ISSN 0887-0624. doi: 10.1021/ef401360k.

[145] B. Magnusson and U. Örnemark. *Eurachem Guide: The Fitness for Purpose of Analytical Methods: A Laboratory Guide to Method Validation and Related Topics*. Eurachem, 2 edition, 2014. ISBN 978-91-87461-59-0.

[146] A. Kruve, R. Rebane, K. Kipper, M.-L. Oldekop, H. Evard, K. Herodes, P. Ravio, and I. Leito. Tutorial review on validation of liquid chromatography-mass spectrometry methods: part II. *Analytica chimica acta*, 870:8–28, 2015. doi: 10.1016/j.aca.2015.02.016.

[147] G. L. Harris. *Selected Laboratory and measurement practices and procedures to support basic mass calibrations*. National Institute of Standards and Technology, Gaithersburg, MD, 2019. doi: 10.6028/NIST.IR.6969-2019.

[148] B. Magnusson, T. Näykki, H. Hovind, M. Krysell, and E. Sahlin. Handbook for calculation of measurement uncertainty in environmental laboratories: Nordtest Report TR 537, 2017.

[149] I. Alkorta and J. Elguero. The carbon–carbon bond dissociation energy as a function of the chain length. *Chemical Physics Letters*, 425(4-6):221–224, 2006. ISSN 00092614. doi: 10.1016/j.cplett.2006.05.050.

[150] K. C. Hunter and A. L. L. East. Properties of C−C Bonds in n-Alkanes: Relevance to Cracking Mechanisms. *The Journal of Physical Chemistry A*, 106(7):1346–1356, 2002. ISSN 1089-5639. doi: 10.1021/jp0129030.

[151] V. D. Knyazev. Effects of chain length on the rates of C-C bond dissociation in linear alkanes and polyethylene. *The journal of physical chemistry. A*, 111(19):3875–3883, 2007. doi: 10.1021/jp066419e.

[152] M. S. Stark, J. J. Wilkinson, J. R. L. Smith, A. Alfadhl, and B. A. Pochopien. Autoxidation of Branched Alkanes in the Liquid Phase. *Industrial & Engineering Chemistry Research*, 50(2):817–823, 2011. ISSN 0888-5885. doi: 10.1021/ie101695g.

[153] J. M. Hudzik, J. W. Bozzelli, and J. M. Simmie. Thermochemistry of C7H16 to C10H22 alkane isomers: primary, secondary, and tertiary C-H bond dissociation energies and effects of branching. *The journal of physical chemistry. A*, 118(40):9364–9379, 2014. doi: 10.1021/jp503587b.

[154] C. Zhu, L. Rui, and Y. Fu. Homolytic Bond Dissociation Enthalpies of C–C and C–H Bonds in Highly Crowded Alkanes. *Chinese Journal of Chemistry*, 26(8):1493–1500, 2008. ISSN 1001-604X. doi: 10.1002/cjoc.200890270.

[155] F. Agapito, P. M. Nunes, B. J. Costa Cabral, R. M. Borges dos Santos, and J. A. Martinho Simões. Energetic differences between the five- and six-membered ring hydrocarbons: strain energies in the parent and radical molecules. *The Journal of organic chemistry*, 73 (16):6213–6223, 2008. doi: 10.1021/jo800690m.

[156] L. Lian, Y. He, L. Xing, C. Xie, Z. Wang, X. Wang, and H. Li. Kinetic study of hydrogen abstraction reactions from n-propyl/n-butylcyclohexane by hydrogen atom. *Fuel*, 354:129348, 2023. ISSN 00162361. doi: 10.1016/j.fuel.2023.129348. URL https://www.sciencedirect.com/science/article/pii/S0016236123019622.

[157] F. Jaffe, T. R. Steadman, and R. W. McKinney. Primary Products of Decalin Autoxidation. *Journal of the American Chemical Society*, 85(3):351–353, 1963. ISSN 0002-7863. doi: 10.1021/ja00886a027.

[158] L. Yue, X. Qin, X. Wu, Y. Guo, L. Xu, H. Xie, and W. Fang. Thermal Decomposition Kinetics and Mechanism of 1,1′-Bicyclohexyl. *Energy & Fuels*, 28(7):4523–4531, 2014. ISSN 0887-0624. doi: 10.1021/ef501077n.

[159] M. Mehl, T. Faravelli, F. Giavazzi, E. Ranzi, P. Scorletti, A. Tardani, and D. Terna. Detailed Chemistry Promotes Understanding of Octane Numbers and Gasoline Sensitivity. *Energy & Fuels*, 20(6):2391–2398, 2006. ISSN 0887-0624. doi: 10.1021/ef060339s.

[160] F. H. Vermeire, R. de Bruycker, O. Herbinet, H.-H. Carstensen, F. Battin-Leclerc, G. B. Marin, and K. M. van Geem. Experimental and kinetic modeling study of the pyrolysis and oxidation of 1,5-hexadiene: The reactivity of allylic radicals and their role in the formation of aromatics. *Fuel*, 208:779–790, 2017. ISSN 00162361. doi: 10.1016/j.fuel.2017.07.042.

[161] C. Huang, P. Zhang, J. Wang, S. Kang, F. Zhang, C. K. Law, and B. Yang. Determination of rate constants for a thermoneutral H-abstraction reaction: Allylic hydrogen abstraction from 1,5-hexadiene by allyl radical. *Proceedings of the Combustion Institute*, 38(1):861–869, 2021. ISSN 15407489. doi: 10.1016/j.proci.2020.07.054.

[162] K. Wang, S. M. Villano, and A. M. Dean. Reactions of allylic radicals that impact molecular weight growth kinetics. *Physical Chemistry Chemical Physics*, 17(9):6255–6273, 2015. ISSN 1463-9084. doi: 10.1039/c4cp05308g.

[163] J. L. Bolland. Kinetics of olefin oxidation. *Quarterly Reviews, Chemical Society*, 3(1):1, 1949. ISSN 0009-2681. doi: 10.1039/qr9490300001.

[164] D. E. van Sickle, F. R. Mayo, R. M. Arluck, and M. G. Syz. Oxidations of Acyclic Alkenes. *Journal of the American Chemical Society*, 89(4):967–977, 1967. ISSN 0002-7863. doi: 10.1021/ja00980a039.

[165] E. G. E. Hawkins and D. C. Quin. Autoxidation of olefins. *Journal of Applied Chemistry*, 6(1):1–11, 1956. ISSN 0021-8871. doi: 10.1002/jctb.5010060101.

[166] Y.-R. Luo, editor. *Comprehensive Handbook of Chemical Bond Energies: BDEs of C–H bonds*. CRC Press, 1 edition, 2007. ISBN 9780429128684.

[167] I. Meziane, N. Delort, O. Herbinet, R. Bounaceur, and F. Battin-Leclerc. A comparative study of the oxidation of toluene and the three isomers of xylene. *Combustion and Flame*, 257:113046, 2023. ISSN 00102180. doi: 10.1016/j.combustflame.2023.113046. URL https://www.sciencedirect.com/science/article/pii/S0010218023004212.

[168] G. Kukkadapu, D. Kang, S. W. Wagnon, K. Zhang, M. Mehl, M. Monge-Palacios, H. Wang, S. S. Goldsborough, C. K. Westbrook, and W. J. Pitz. Kinetic modeling study of surrogate components for gasoline, jet and diesel fuels: C7-C11 methylated aromatics. *Proceedings of the Combustion Institute*, 37(1):521–529, 2019. ISSN 15407489. doi: 10.1016/j.proci.2018.08.016. URL https://www.sciencedirect.com/science/article/pii/S1540748918305583.

[169] C. R. Shaddix, K. Brezinsky, and I. Glassman. Oxidation of 1-methylnaphthalene. *Symposium (International) on Combustion*, 24(1):683–690, 1992. ISSN 00820784. doi: 10.1016/S0082-0784(06)80084-6. URL https://www.sciencedirect.com/science/article/pii/S0082078406800846.

[170] W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle, and E. V. Anslyn. The Evolution of Data-Driven Modeling in Organic Chemistry. *ACS central science*, 7(10): 1622–1637, 2021. ISSN 2374-7943. doi: 10.1021/acscentsci.1c00535.

[171] F. K. Brown. Chemoinformatics: What is it and How does it Impact Drug Discovery. volume 33 of *Annual Reports in Medicinal Chemistry*, pages 375–384. Elsevier, 1998. ISBN 9780120405336. doi: 10.1016/S0065-7743(08)61100-8.

[172] A. Varnek and I. I. Baskin. Chemoinformatics as a Theoretical Chemistry Discipline. *Molecular informatics*, 30(1):20–32, 2011. ISSN 1868-1743. doi: 10.1002/minf.201000100.

[173] K. Roy, S. Kar, and R. N. Das. QSAR/QSPR Modeling: Introduction. In K. Roy, S. Kar, and R. N. Das, editors, *A primer on QSARQSPR modeling*, SpringerBriefs in Molecular Science, pages 1–36. Springer Berlin Heidelberg, New York NY, 2015. ISBN 978-3-319-17280-4. doi: 10.1007/978-3-319-17281-1{\textunderscore}1.

[174] S. Yousefinejad and B. Hemmateenejad. Chemometrics tools in QSAR/QSPR studies: A historical perspective. *Chemometrics and Intelligent Laboratory Systems*, 149:177–204, 2015. ISSN 01697439. doi: 10.1016/j.chemolab.2015.06.016.

[175] D. A. Saldana, B. Creton, P. Mougin, N. Jeuland, B. Rousseau, and L. Starck. Rational Formulation of Alternative Fuels using QSPR Methods: Application to Jet Fuels. *Oil & Gas Science and Technology – Revue d'IFP Energies nouvelles*, 68(4):651–662, 2013. ISSN 1294-4475. doi: 10.2516/ogst/2012034.

[176] R. Li, J. M. Herreros, A. Tsolakis, and W. Yang. Machine learning-quantitative structure property relationship (ML-QSPR) method for fuel physicochemical properties prediction of multiple fuel types. *Fuel*, 304:121437, 2021. ISSN 00162361. doi: 10.1016/j.fuel.2021.121437.

[177] M. Karelson, V. S. Lobanov, and A. R. Katritzky. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chemical reviews*, 96(3):1027–1044, 1996. doi: 10.1021/cr950202r.

[178] G. M. Maggiora. On Outliers And Activity Cliffs-Why QSAR Often Disappoints. *Journal of chemical information and modeling*, 46(4):1535, 2006. ISSN 1549-9596. doi: 10.1021/ci060117s.

[179] Aviation Fuels: Technical Review. URL https://www.chevron.com/-/media/chevron/operations/documents/aviation-tech-review.pdf.

[180] L. Pattanaik and C. W. Coley. Molecular Representation: Going Long on Fingerprints. *Chem*, 6(6):1204–1207, 2020. ISSN 24519294. doi: 10.1016/j.chempr.2020.05.002.

[181] R. Todeschini, V. Consonni, and P. Gramatica. Chemometrics in QSAR. In S. D. Brown, editor, *Comprehensive chemometrics*, pages 129–172. Elsevier, Amsterdam, 2009. ISBN 9780444527011. doi: 10.1016/B978-044452701-1.00007-7.

[182] S. Raghunathan and U. D. Priyakumar. Molecular representations for machine learning applications in chemistry. *International Journal of Quantum Chemistry*, 122(7), 2022. ISSN 0020-7608. doi: 10.1002/qua.26870.

[183] Daylight Chemical Information Systems, Inc. SMARTS, Daylight Theory Manual. URL https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

[184] B. Sanchez-Lengeling, E. Reif, A. Pearce, and A. Wiltschko. A Gentle Introduction to Graph Neural Networks. *Distill*, 6(8), 2021. ISSN 2476-0757. doi: 10.23915/distill.00033.

[185] J. Yang, Y. Cai, K. Zhao, H. Xie, and X. Chen. Concepts and applications of chemical fingerprint for hit and lead screening. *Drug discovery today*, 27(11):103356, 2022. doi: 10.1016/j.drudis.2022.103356. URL https://www.sciencedirect.com/science/article/pii/S135964462200349X.

[186] F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, and F. Glorius. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem*, 6(6):1379–1390, 2020. ISSN 24519294. doi: 10.1016/j.chempr.2020.02.017.

[187] Danishuddin and A. U. Khan. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug discovery today*, 21(8):1291–1302, 2016. doi: 10.1016/j.drudis.2016.06.013.

[188] R. Li, J. M. Herreros, A. Tsolakis, and W. Yang. Integrated machine learning-quantitative structure property relationship (ML-QSPR) and chemical kinetics for high throughput fuel screening toward internal combustion engine. *Fuel*, 307:121908, 2022. ISSN 00162361. doi: 10.1016/j.fuel.2021.121908.

[189] C. Nieto-Draghi, G. Fayet, B. Creton, X. Rozanska, P. Rotureau, J.-C. de Hemptinne, P. Ungerer, B. Rousseau, and C. Adamo. A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes. *Chemical reviews*, 115(24):13093–13164, 2015. doi: 10.1021/acs.chemrev.5b00215.

[190] J. S. Duca and A. J. Hopfinger. Estimation of molecular similarity based on 4D-QSAR analysis: formalism and validation. *Journal of chemical information and computer sciences*, 41(5):1367–1387, 2001. ISSN 0095-2338. doi: 10.1021/ci0100090.

[191] C. H. Andrade, K. F. M. Pasqualoto, E. I. Ferreira, and A. J. Hopfinger. 4D-QSAR: perspectives in drug design. *Molecules*, 15(5):3281–3294, 2010. doi: 10.3390/molecules15053281.

[192] F. Grisoni, D. Ballabio, R. Todeschini, and V. Consonni. Molecular Descriptors for Structure-Activity Applications: A Hands-On Approach. *Methods in molecular biology (Clifton, N.J.)*, 1800:3–53, 2018. doi: 10.1007/978-1-4939-7899-1{\textunderscore}1.

[193] D. A. Saldana, L. Starck, P. Mougin, B. Rousseau, L. Pidol, N. Jeuland, and B. Creton. Flash Point and Cetane Number Predictions for Fuel Compounds Using Quantitative Structure Property Relationship (QSPR) Methods. *Energy & Fuels*, 25(9):3900–3908, 2011. ISSN 0887-0624. doi: 10.1021/ef200795j.

[194] B. Creton, B. Veyrat, and M.-H. Klopffer. Fuel sorption into polymers: Experimental and machine learning studies. *Fluid Phase Equilibria*, 556:113403, 2022. ISSN 03783812. doi: 10.1016/j.fluid.2022.113403.

[195] R. Li, J. M. Herreros, A. Tsolakis, and W. Yang. Novel Functional Group Contribution Method for Surrogate Formulation with Accurate Fuel Compositions. *Energy & Fuels*, 34 (3):2989–3012, 2020. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.9b04270.

[196] R. Li, J. M. Herreros, A. Tsolakis, and W. Yang. Machine learning regression based group contribution method for cetane and octane numbers prediction of pure fuel compounds and mixtures. *Fuel*, 280:118589, 2020. ISSN 00162361. doi: 10.1016/j.fuel.2020.118589.

[197] R. Li, J. M. Herreros, A. Tsolakis, and W. Yang. Machine learning and deep learning enabled fuel sooting tendency prediction from molecular structure. *Journal of molecular graphics & modelling*, 111:108083, 2022. doi: 10.1016/j.jmgm.2021.108083.

[198] S. Ajmani, S. C. Rogers, M. H. Barley, and D. J. Livingstone. Application of QSPR to mixtures. *Journal of chemical information and modeling*, 46(5):2043–2055, 2006. ISSN 1549-9596. doi: 10.1021/ci050559o.

[199] T. Gaudin, P. Rotureau, and G. Fayet. Combining mixing rules with QSPR models for pure chemicals to predict the flash points of binary organic liquid mixtures. *Fire Safety Journal*, 74:61–70, 2015. ISSN 03797112. doi: 10.1016/j.firesaf.2015.04.006.

[200] C. Hall, B. Creton, B. Rauch, U. Bauder, and M. Aigner. Probabilistic Mean Quantitative Structure–Property Relationship Modeling of Jet Fuel Properties. *Energy & Fuels*, 36(1): 463–479, 2022. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.1c03334.

[201] R. P. Sheridan. The centroid approximation for mixtures: calculating similarity and deriving structure–activity relationships. *Journal of chemical information and computer sciences*, 40(6):1456–1469, 2000. ISSN 0095-2338. doi: 10.1021/ci000045j.

[202] X. Shi, H. Li, Z. Song, X. Zhang, and G. Liu. Quantitative composition-property relationship of aviation hydrocarbon fuel based on comprehensive two-dimensional gas chromatography with mass spectrometry and flame ionization detector. *Fuel*, 200: 395–406, 2017. ISSN 00162361. doi: 10.1016/j.fuel.2017.03.073.

[203] J. R. Koza, F. H. Bennett, D. Andre, and M. A. Keane. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In J. S. Gero and F. Sudweeks, editors, *Artificial Intelligence in Design '96*, pages 151–170. Springer Netherlands, Dordrecht, 1996. ISBN 978-94-010-6610-5. doi: 10.1007/978-94-009-0279-4{\textunderscore}9.

[204] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science (New York, N.Y.)*, 349(6245):255–260, 2015. doi: 10.1126/science.aaa8415.

[205] W. Wang. Supervised Learning Paradigm. In W. Wang, editor, *Principles of Machine Learning*, pages 269–290. Springer Nature Singapore, Singapore, 2025. ISBN 978-981-97-5332-1. doi: 10.1007/978-981-97-5333-8{\textunderscore}8.

[206] Y.-m. Dai, Z.-p. Zhu, Z. Cao, Y.-f. Zhang, J.-l. Zeng, and X. Li. Prediction of boiling points of organic compounds by QSPR tools. *Journal of molecular graphics & modelling*, 44: 113–119, 2013. doi: 10.1016/j.jmgm.2013.04.007.

[207] A. Lavecchia. Machine-learning approaches in drug discovery: methods and applications. *Drug discovery today*, 20(3):318–331, 2015. doi: 10.1016/j.drudis.2014.10.012.

[208] R. Xu and D. Wunsch. Survey of clustering algorithms. *IEEE transactions on neural networks*, 16(3):645–678, 2005. ISSN 1045-9227. doi: 10.1109/TNN.2005.845141.

[209] R. Bro and A. K. Smilde. Principal component analysis. *Anal. Methods*, 6(9):2812–2831, 2014. ISSN 1759-9660. doi: 10.1039/C3AY41907J.

[210] L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[211] B. Sridharan, A. Sinha, J. Bardhan, R. Modee, M. Ehara, and U. D. Priyakumar. Deep reinforcement learning in chemistry: A review. *Journal of Computational Chemistry*, 45 (22):1886–1898, 2024. ISSN 0192-8651. doi: 10.1002/jcc.27354.

[212] J. Starmer. *The StatQuest illustrated guide to machine learning!!! Triple bam!!!* Josh Starmer, USA?, 2022. ISBN 9798986924007.

[213] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. ISSN 0885-6125. doi: 10.1007/BF00994018.

[214] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324.

[215] T. Hastie, R. Tibshirani, and J. Friedman. Boosting and Additive Trees. In T. Hastie, R. Tibshirani, and J. H. Friedman, editors, *The elements of statistical learning*, Springer series in statistics, pages 337–387. Springer, New York NY, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7{\textunderscore}10.

[216] S. C. Peter, J. K. Dhanjal, V. Malik, N. Radhakrishnan, M. Jayakanthan, and D. Sundar. Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications. In *Encyclopedia of Bioinformatics and Computational - Ranganathan, Gribskov et al. (Ed.) 2019 –*, pages 661–676. doi: 10.1016/B978-0-12-809633-8.20197-0. URL https://www.sciencedirect.com/science/article/pii/B9780128096338201970.

[217] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional Networks on Graphs for Learning Molecular Fingerprints.

[218] L. Yang and A. Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020. ISSN 0925-2312. doi: 10.1016/j.neucom.2020.07.061. URL https://www.sciencedirect.com/science/article/pii/S0925231220311693.

[219] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13(null):281–305, 2012. ISSN 1532-4435.

[220] V. Nguyen. Bayesian Optimization for Accelerating Hyper-Parameter Tuning. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 302–305, 2019. doi: 10.1109/AIKE.2019.00060.

[221] T. Hastie, R. Tibshirani, and J. Friedman. Overview of Supervised Learning. In *The Elements of Statistical Learning*, pages 9–41. doi: 10.1007/978-0-387-84858-7{\textunderscore}2.

[222] M. Kuhn and K. Johnson. Over-Fitting and Model Tuning. In *Applied Predictive Modeling*, pages 61–80. doi: 10.1007/978-1-4614-6849-3.

[223] F. Westad and F. Marini. Validation of chemometric models - a tutorial. *Analytica chimica acta*, 893:14–24, 2015. doi: 10.1016/j.aca.2015.06.056.

[224] L. Xu, Q.-S. Xu, M. Yang, H.-Z. Zhang, C.-B. Cai, J.-H. Jiang, H.-L. Wu, and R.-Q. Yu. On estimating model complexity and prediction errors in multivariate calibration: generalized resampling by random sample weighting (RSW). *Journal of Chemometrics*, 25(2):51–58, 2011. ISSN 08869383. doi: 10.1002/cem.1323.

[225] J. D. Rodríguez, A. Pérez, and J. A. Lozano. Sensitivity analysis of kappa-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):569–575, 2010. doi: 10.1109/TPAMI.2009.187.

[226] S. Aminane, M. Sicard, Y. Melliti, B. Raepsaet, L. Sicard, and F. Ser. Highlighting the "structure–reactivity" relationship and the impact on the kinetics for the autoxidation reaction of hydrocarbons. *Fuel*, 370:131748, 2024. ISSN 00162361. doi: 10.1016/j.fuel.2024.131748.

[227] G. Landrum, P. Tosco, B. Kelley, Ric, sriniker, gedeck, R. Vianello, D. Cosgrove, NadineSchneider, E. Kawashima, D. N, A. Dalke, G. Jones, B. Cole, M. Swain, S. Turk, AlexanderSavelyev, A. Vaucher, M. Wójcikowski, I. Take, D. Probst, V. F. Scalfani, K. Ujihara, g. godin, A. Pahl, F. Berenger, JLVarjo, jasondbiggs, strets123, and JP. RDKit: Open-source cheminformatics, 2022. URL https://www.rdkit.org.

[228] A. R. Katritzky, M. Kuanar, S. Slavov, C. D. Hall, M. Karelson, I. Kahn, and D. A. Dobchev. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chemical reviews*, 110(10):5714–5789, 2010. doi: 10.1021/cr900238d.

[229] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, New York, 1992. Association for Computing Machinery. ISBN 089791497X. doi: 10.1145/130385.130401.

[230] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 11 of *KDD '16*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

[231] R Core Team. R: A Language and Environment for Statistical Computing, 2021. URL https://www.R-project.org/.

[232] M. Kuhn. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, Articles*, 28(5):1–26, 2008. ISSN 1548-7660. doi: 10.18637/jss.v028.i05. URL https://www.jstatsoft.org/v028/i05.

[233] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis. kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004. doi: 10.18637/jss.v011.i09.

[234] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, Y. Li, and J. Yuan. xgboost: Extreme Gradient Boosting, 2024. URL https://CRAN.R-project.org/package=xgboost.

[235] J.-P. Vert, B. Schölkopf, and K. Tsuda. A Primer on Kernel Methods. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel methods in computational biology*, Computational biology. MIT Press, Cambridge, Mass., 2004. ISBN 9780262256926. doi: 10.7551/mitpress/4057.003.0004.

[236] R. Moreno Jimenez, B. Creton, and S. Marre. Machine learning-based models for accessing thermal conductivity of liquids at different temperature conditions. *SAR and QSAR in environmental research*, 34(8):605–617, 2023. doi: 10.1080/1062936X.2023.2244410.

[237] R. W. Kennard and L. A. Stone. Computer Aided Design of Experiments. *Technometrics*, 11(1):137, 1969. ISSN 00401706. doi: 10.2307/1266770.

[238] Q. Wu, C. Marina-Montes, J. O. Cáceres, J. Anzano, V. Motto-Ros, and L. Duponchel. Interesting features finder (IFF): Another way to explore spectroscopic imaging data sets giving minor compounds and traces a chance to express themselves. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 195:106508, 2022. ISSN 0584-8547. doi: 10.1016/j.sab.2022.106508. URL https://www.sciencedirect.com/science/article/pii/S0584854722001525.

[239] B. Creton, E. Barraud, and C. Nieto-Draghi. Prediction of critical micelle concentration for per- and polyfluoroalkyl substances. *SAR and QSAR in environmental research*, 35(4):309–324, 2024. doi: 10.1080/1062936X.2024.2337011.

[240] B. Creton, N. Brassart, A. Herbaut, and M. Matrat. Numerical Approaches to Determine Cetane Number of Hydrocarbons and Oxygenated Compounds, Mixtures, and their Blends. *Energy & Fuels*, 38(16):15652–15661, 2024. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.4c03007.

[241] I. Cortes-Ciriano and A. Bender. Improved Chemical Structure-Activity Modeling Through Data Augmentation. *Journal of chemical information and modeling*, 55(12): 2682–2692, 2015. ISSN 1549-9596. doi: 10.1021/acs.jcim.5b00570.

[242] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, and S. Stumpf. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024. ISSN 15662535. doi: 10.1016/j.inffus.2024.102301.

[243] S. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. URL http://arxiv.org/pdf/1705.07874.

[244] L. S. Shapley. A Value for n-Person Games. *Contributions to the Theory of Games, Volume II*, pages 307–318, 1953. ISSN 97814008. URL https://www.degruyter.com/document/doi/10.1515/9781400881970-018/html.

[245] K. Aas, M. Jullum, and A. Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298: 103502, 2021. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103502.

[246] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. URL http://arxiv.org/pdf/1802.03888.

[247] OEIS Foundation Inc. A000602: Number of n-node unrooted quartic trees; number of n-carbon alkanes C(n)H(2n+2) ignoring stereoisomers, 10/09/2024. URL https://oeis.org/A000602.

[248] A. Venegas-Reynoso, L. Giarracca-Mehl, M. Lacoue-Negre, B. Creton, C. Ruckebusch, and L. Duponchel. Identification of Key Molecular Features in Liquid Phase Autoxidation of Hydrocarbons. *Energy & Fuels*, 39(2):1192–1201, 2025. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.4c04653.

[249] O. Martinez, K. N. Crabtree, C. A. Gottlieb, J. F. Stanton, and M. C. McCarthy. An accurate molecular structure of phenyl, the simplest aryl radical. *Angewandte Chemie (International ed. in English)*, 54(6):1808–1811, 2015. doi: 10.1002/anie.201409896.

[250] LibreTexts. Emission and Absorbance Spectra, 2022. URL https://chem.libretexts.org/@go/page/365610.

[251] P. Atkins and J. de Paula. *Physical Chemistry*. W.H. Freeman & Company, 8 edition, 2006.

[252] R. S. Krishnan and R. K. Shankar. Raman effect: History of the discovery. *Journal of Raman Spectroscopy*, 10(1):1–8, 1981. ISSN 1097-4555. doi: 10.1002/jrs.1250100103.

[253] LibreTexts. The Electromagnetic Spectrum, 2023. URL https://chem.libretexts.org/@go/page/210898.

[254] M. L.C. Passos and M.F.S. Saraiva, M. Lúcia. Detection in UV-visible spectrophotometry: Detectors, detection systems, and detection strategies. *Measurement*, 135:896–904, 2019. ISSN 0263-2241. doi: 10.1016/j.measurement.2018.12.045.

[255] LibreTexts. Electronic Spectra: Ultraviolet and Visible Spectroscopy, 2015. URL https://chem.libretexts.org/@go/page/32468.

[256] LibreTexts. The Beer-Lambert Law, 2023. URL https://chem.libretexts.org/@go/page/3747.

[257] R. R. de Oliveira, K. M. G. de Lima, R. Tauler, and A. de Juan. Application of correlation constrained multivariate curve resolution alternating least-squares methods for determination of compounds of interest in biodiesel blends using NIR and UV-visible spectroscopic data. *Talanta*, 125:233–241, 2014. doi: 10.1016/j.talanta.2014.02.073.

[258] LibreTexts. Infrared Spectroscopy, 2023. URL https://chem.libretexts.org/@go/page/1847.

[259] J. Workman and L. Weyer. *Practical guide and spectral atlas for interpretive near-infrared spectroscopy, second edition*. CRC Press, Boca Raton, FL, 2 edition, 2012.

[260] LibreTexts. Combination Bands, Overtones and Fermi Resonances, 2023. URL https://chem.libretexts.org/@go/page/1853.

[261] LibreTexts. Near-Infrared and Far-Infrared Spectroscopy, 2022. URL https://chem.libretexts.org/@go/page/379844.

[262] C. Pasquini. Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society*, 14(2):198–219, 2003. doi: 10.1590/s0103-50532003000200006.

[263] R. Velvarská, A. Vráblík, M. Fiedlerová, and R. Černý. Near-infrared spectroscopy for determining the oxidation stability of diesel, biodiesel and their mixtures. *Chemical Papers*, 73(12):2987–2993, 2019. ISSN 2585-7290. doi: 10.1007/s11696-019-00852-4.

[264] S. Wang, S. Liu, Y. Yuan, J. Zhang, Z. Wang, and X. Che. A novel CC-tSNE-SVR model for rapid determination of diesel fuel quality by near infrared spectroscopy. *Infrared Physics & Technology*, 106:103276, 2020. ISSN 13504495. doi: 10.1016/j.infrared.2020.103276.

[265] G. Varghese, K. Saeed, and K. J. Rutt. Determination of the oxidative stability of biodiesel fuels by near-infrared spectroscopy. *Fuel*, 290:120015, 2021. ISSN 00162361. doi: 10.1016/j.fuel.2020.120015.

[266] L. A. Carbognani Ortega, J. K. Rodriguez Guerrero, E. S. Ferreira, J. C. Arambarri, M. Bartolini, and P. R. Pereira-Almao. Infrared spectroscopy for carboxylic acid and phenol determination in biocrude and its derived products. *Sustainable Energy & Fuels*, 4(3):1157–1167, 2020. doi: 10.1039/C9SE00376B.

[267] Li Jingyan, C. Xiaoli, and T. Songbai. Research on Determination of Total Acid Number of Petroleum Using Mid-infrared Attenuated Total Reflection Spectroscopy. *Energy & Fuels*, 26(9):5633–5637, 2012. ISSN 0887-0624. doi: 10.1021/ef3002372.

[268] Y. Wen, S. Zhou, L. Wang, Q. Li, Y. Gao, and X. Yu. New Method for the Determination of the Induction Period of Walnut Oil by Fourier Transform Infrared Spectroscopy. *Food Analytical Methods*, 15(3):833–843, 2022. ISSN 1936-9751. doi: 10.1007/s12161-021-02170-6.

[269] P. Rostron and D. Gerber. Raman Spectroscopy, a review. *International Journal of Engineering and Technical Research*, 6:50–64, 2016.

[270] LibreTexts. Raman Spectroscopy, 2022. URL https://chem.libretexts.org/@go/page/148443.

[271] Z. Liu, N. Luo, J. Shi, Y. Zhang, C. Xie, W. Zhang, H. Wang, X. He, and Z. Chen. Raman spectroscopy for the discrimination and quantification of fuel blends. *Journal of Raman Spectroscopy*, 50(7):1008–1014, 2019. ISSN 1097-4555. doi: 10.1002/jrs.5602.

[272] J. B. Cooper, K. L. Wise, J. Groves, and W. T. Welch. Determination of octane numbers and Reid vapor pressure of commercial petroleum fuels using FT-Raman spectroscopy and partial least-squares regression analysis. *Analytical chemistry*, 67(22):4096–4100, 1995.

[273] D. Ambre, M. Sheyyab, P. Lynch, E. K. Mayhew, and K. Brezinsky. A Raman spectroscopy based chemometric approach to predict the derived cetane number of hydrocarbon jet fuels and their mixtures. *Talanta*, 271:125635, 2024. doi: 10.1016/j.talanta.2024.125635. URL https://www.sciencedirect.com/science/article/pii/S0039914024000146.

[274] LibreTexts. Introduction to NMR, 2015. URL https://chem.libretexts.org/@go/page/1834.

[275] P. M. V. Raja and A. R. Barron. NMR Spectroscopy, 2022. URL https://chem.libretexts.org/@go/page/55887.

[276] R. Kumar, V. Bansal, M. B. Patel, and A. S. Sarpal. 1 H Nuclear Magnetic Resonance (NMR) Determination of the Iodine Value in Biodiesel Produced from Algal and Vegetable Oils. *Energy & Fuels*, 26(11):7005–7008, 2012. ISSN 0887-0624. doi: 10.1021/ef300991n.

[277] LibreTexts. EPR - Introduction, 2023. URL https://chem.libretexts.org/@go/page/1793.

[278] LibreTexts. EPR - Theory, 2023. URL https://chem.libretexts.org/@go/page/1794.

[279] LibreTexts. EPR - Interpretation, 2023. URL https://chem.libretexts.org/@go/page/1792.

[280] M. L. Andersen. Chapter 10 - Lipid oxidation studied by electron paramagnetic resonance (EPR). In P. J. García-Moreno, C. Jacobsen, A.-D. M. Sørensen, and B. Yesiltas, editors, *Omega-3 Delivery Systems*, pages 201–213. Academic Press, 2021. ISBN 978-0-12-821391-9. doi: 10.1016/B978-0-12-821391-9.00004-1. URL https://www.sciencedirect.com/science/article/pii/B9780128213919000041.

[281] N. Babić, S. Pondaven, and H. Vezin. EPR Spin-Trapping Study of Free Radical Intermediates in Polyalphaolefin Base Oil Autoxidation. *Polymer Degradation and Stability*, 192:109687, 2021. ISSN 0141-3910. doi: 10.1016/j.polymdegradstab.2021.109687. URL https://www.sciencedirect.com/science/article/pii/S014139102100207X.

[282] S. Wold. Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, 30(1):109–115, 1995. ISSN 01697439. doi: 10.1016/0169-7439(95)00042-9.

[283] K. Héberger. Chemoinformatics—multivariate mathematical–statistical methods for data evaluation. In K. Vékey, A. Telekes, and A. Vertes, editors, *Medical applications of mass spectrometry*, pages 141–169. Elsevier, Amsterdam and Boston, 2008. ISBN 9780444519801. doi: 10.1016/B978-044451980-1.50009-4.

[284] S. D. Brown. The chemometrics revolution re-examined. *Journal of Chemometrics*, 31(1): e2856, 2017. ISSN 08869383. doi: 10.1002/cem.2856.

[285] R. Bro and A. K. Smilde. Centering and scaling in component analysis. *Journal of Chemometrics*, 17(1):16–33, 2003. ISSN 08869383. doi: 10.1002/cem.773.

[286] K. H. Liland, T. Almøy, and B.-H. Mevik. Optimal choice of baseline correction for multivariate calibration of spectra. *Applied spectroscopy*, 64(9):1007–1016, 2010. doi: 10.1366/000370210792434350.

[287] J. Engel, J. Gerretzen, E. Szymańska, J. J. Jansen, G. Downey, L. Blanchet, and L. M. Buydens. Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, 50:96–106, 2013. ISSN 01659936. doi: 10.1016/j.trac.2013.04.015.

[288] T. Fearn, C. Riccioli, A. Garrido-Varo, and J. E. Guerrero-Ginel. On the geometry of SNV and MSC. *Chemometrics and Intelligent Laboratory Systems*, 96(1):22–26, 2009. ISSN 01697439. doi: 10.1016/j.chemolab.2008.11.006.

[289] M. S. Dhanoa, S. J. Lister, R. Sanderson, and R. J. Barnes. The Link between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) Transformations of NIR Spectra. *J. Near Infrared Spectrosc.*, 2(1):43–47, 1994.

[290] A. Savitzky and M. J. E. Golay. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964. ISSN 0003-2700. doi: 10.1021/ac60214a047.

[291] A. Davies and T. Fearn. Back to basics: Removing multiplicative effects (1). *Spectroscopy Europe*, 19:24–28, 2007.

[292] J. Peng, S. Peng, Jiang, J. Wei, C. Li, and J. Tan. Asymmetric least squares for multiple spectra baseline correction. *Analytica chimica acta*, 683(1):63–68, 2010. doi: 10.1016/j.aca.2010.08.033.

[293] M. D. Peris-Díaz and A. Krężel. A guide to good practice in chemometric methods for vibrational spectroscopy, electrochemistry, and hyphenated mass spectrometry. *TrAC Trends in Analytical Chemistry*, 135:116157, 2021. ISSN 01659936. doi: 10.1016/j.trac.2020.116157.

[294] M. X. Cohen. *Linear Algebra: Theory, Intuiton, Code.* Sincxpress Bv, 2021.

[295] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065):20150202, 2016. doi: 10.1098/rsta.2015.0202.

[296] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. URL https://arxiv.org/pdf/1802.03426.

[297] D. Kobak and G. C. Linderman. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature biotechnology*, 39(2):156–157, 2021. doi: 10.1038/s41587-020-00809-z.

[298] M. Joswiak, Y. Peng, I. Castillo, and L. H. Chiang. Dimensionality reduction for visualizing industrial chemical process data. *Control Engineering Practice*, 93:104189, 2019. ISSN 0967-0661. doi: 10.1016/j.conengprac.2019.104189.

[299] C. Phechkrajang, P. Khongkaew, W. Limwikrant, and M. Jaturanpinyo. Non-Destructive Analysis of Chlorpheniramine Maleate Tablets and Granules by Chemometrics-Assisted Attenuated Total Reflectance Infrared Spectroscopy. *Molecules*, 27(12):3760, 2022. doi: 10.3390/molecules27123760.

[300] T. Fearn. Standardisation and Calibration Transfer for near Infrared Instruments: A Review. *Journal of Near Infrared Spectroscopy*, 9(4):229–244, 2001. ISSN 1751-6552. doi: 10.1255/jnirs.309.

[301] S. Wold, M. Sjöström, and L. Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001. ISSN 01697439. doi: 10.1016/S0169-7439(01)00155-1.

[302] S. de Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993. ISSN 01697439. doi: 10.1016/0169-7439(93)85002-X.

[303] J. A. Cayuela Sánchez, W. Moreda, and J. M. García. Rapid determination of olive oil oxidative stability and its major quality parameters using vis/NIR transmittance spectroscopy. *Journal of agricultural and food chemistry*, 61(34):8056–8062, 2013. doi: 10.1021/jf4021575.

[304] B. Ustün, W. J. Melssen, and L. M. C. Buydens. Visualisation and interpretation of Support Vector Regression models. *Analytica chimica acta*, 595(1-2):299–309, 2007. doi: 10.1016/j.aca.2007.03.023. URL https://www.sciencedirect.com/science/article/pii/S0003267007004904.

[305] Å. Björck. *Numerical methods for least squares problems*. SIAM, Philadelphia, MS, 1996.

[306] K. H. Liland, T. Almøy, and B.-H. Mevik. Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra. *Applied spectroscopy*, 64:1007–1016, 2010. doi: 10.1366/000370210792434350.

[307] C. Beleites and V. Sergo. hyperSpec: a package to handle hyperspectral data sets in R, 0.100.2. URL https://github.com/r-hyperspec/hyperSpec.

[308] O. H. Wheeler. Near Infrared Spectra Of Organic Compounds. *Chemical reviews*, 59(4): 629–666, 1959. doi: 10.1021/cr50028a004.

[309] Y. Binev, M. M. B. Marques, and J. Aires-de Sousa. Prediction of 1H NMR Coupling Constants with Associative Neural Networks Trained for Chemical Shifts. *Journal of chemical information and modeling*, 47(6):2089–2097, 2007. ISSN 1549-9596. doi: 10.1021/ci700172n.

[310] A. M. Castillo, L. Patiny, and J. Wist. Fast and accurate algorithm for the simulation of NMR spectra of large spin systems. *Journal of Magnetic Resonance*, 209(2):123–130, 2011. ISSN 1090-7807. doi: 10.1016/j.jmr.2010.12.008. URL https://www.sciencedirect.com/science/article/pii/S1090780710004003.

[311] A. Leniak, W. Pietruś, and R. Kurczab. From NMR to AI: Designing a Novel Chemical Representation to Enhance Machine Learning Predictions of Physicochemical Properties. *Journal of chemical information and modeling*, 64(8):3302–3321, 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c02039.

[312] P. Willett. Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *Journal of computational biology : a journal of computational molecular cell biology*, 6(3-4):447–457, 1999. ISSN 1066-5277. doi: 10.1089/106652799318382.

Appendix A

List of the hydrocarbons in the jet-fuel range used in this work.

Table A.1 – List of commercially available hydrocarbons.

Compound	Hydrocarbon family	Purity (%)	Provider
n-pentane	*n*-paraffins (C_nH_{2n+2})	98	Fisher Scientific
n-hexane	*n*-paraffins (C_nH_{2n+2})	99	Fisher Scientific
n-heptane	*n*-paraffins (C_nH_{2n+2})	99	Fisher Scientific
n-octane	*n*-paraffins (C_nH_{2n+2})	>99	Fisher Scientific
n-nonane	*n*-paraffins (C_nH_{2n+2})	99	Fisher Scientific
n-decane	*n*-paraffins (C_nH_{2n+2})	>99	Fisher Scientific
n-undecane	*n*-paraffins (C_nH_{2n+2})	99	Fisher Scientific
n-dodecane	*n*-paraffins (C_nH_{2n+2})	99	Fisher Scientific
n-tridecane	*n*-paraffins (C_nH_{2n+2})	>99	Fisher Scientific
n-tetradecane	*n*-paraffins (C_nH_{2n+2})	>99	Fisher Scientific
n-pentadecane	*n*-paraffins (C_nH_{2n+2})	99	Fisher Scientific
n-hexadecane	*n*-paraffins (C_nH_{2n+2})	99	Fisher Scientific
n-heptadecane	*n*-paraffins (C_nH_{2n+2})	99	Fisher Scientific
n-octadecane	*n*-paraffins (C_nH_{2n+2})	>99	Fisher Scientific
n-nonadecane	*n*-paraffins (C_nH_{2n+2})	99	Fisher Scientific
n-eicosane	*n*-paraffins (C_nH_{2n+2})	99	Fisher Scientific
2-methylpentane	iso-paraffins (C_nH_{2n+2})	>99	Fisher Scientific
2,2-dimethylbutane	iso-paraffins (C_nH_{2n+2})	>98	Merck
3-methylpentane	iso-paraffins (C_nH_{2n+2})	>99	Fisher Scientific
2,3-dimethylbutane	iso-paraffins (C_nH_{2n+2})	99	Fisher Scientific
2-methylheptane	iso-paraffins (C_nH_{2n+2})	99	Fisher Scientific
2,5-dimethylhexane	iso-paraffins (C_nH_{2n+2})	99	Fisher Scientific
2,2,4-trimethylpentane	iso-paraffins (C_nH_{2n+2})	> 99	Fisher Scientific
2,3,4-trimethylpentane	iso-paraffins (C_nH_{2n+2})	>98	Fisher Scientific
2-methylnonane	iso-paraffins (C_nH_{2n+2})	98	Fisher Scientific
2,2,4,6,6-pentamethylheptane	iso-paraffins (C_nH_{2n+2})	99	Fisher Scientific
isododecane, isomer mixture	iso-paraffins (C_nH_{2n+2})	NA	Fisher Scientific
2,2,4,4,6,8,8-heptamethylnonane	iso-paraffins (C_nH_{2n+2})	98	Fisher Scientific
cyclopentane	mono-naphthenes (C_nH_{2n})	97	Fisher Scientific
cyclohexane	mono-naphthenes (C_nH_{2n})	>99	Fisher Scientific
methylcyclohexane	mono-naphthenes (C_nH_{2n})	99	Fisher Scientific
cycloheptane	mono-naphthenes (C_nH_{2n})	99	Fisher Scientific
cyclooctane	mono-naphthenes (C_nH_{2n})	>99	Fisher Scientific

Continued on next page

Table A.1 – continued from previous page.

Compound	Family	Purity (%)	Provider
ethylcyclohexane	mono-naphthenes (C_nH_{2n})	>99	Fisher Scientific
1,3-dimethylcyclohexane, mixture of *cis* and *trans*	mono-naphthenes (C_nH_{2n})	99	Merck
1,2,4-trimethylcyclohexane	mono-naphthenes (C_nH_{2n})	97	Merck
tert-butylcyclohexane	mono-naphthenes (C_nH_{2n})	>99	Fisher Scientific
n-butylcyclohexane	mono-naphthenes (C_nH_{2n})	99	Fisher Scientific
1-hexene	mono-olefins (C_nH_{2n})	99	Fisher Scientific
trans-2-hexene	mono-olefins (C_nH_{2n})	>99	Fisher Scientific
3,3-dimethyl-1-butene	mono-olefins (C_nH_{2n})	95	Fisher Scientific
2,3-dimethyl-1-butene	mono-olefins (C_nH_{2n})	99	Fisher Scientific
2,3-dimethyl-2-butene	mono-olefins (C_nH_{2n})	98	Fisher Scientific
1-octene	mono-olefins (C_nH_{2n})	>99	Fisher Scientific
1-dodecene	mono-olefins (C_nH_{2n})	96	Fisher Scientific
1-hexadecene	mono-olefins (C_nH_{2n})	94	Fisher Scientific
2,4,4-trimethyl-1-pentene	mono-olefins (C_nH_{2n})	99	Fisher Scientific
2,4,4-trimethyl-2-pentene	mono-olefins (C_nH_{2n})	99	Fisher Scientific
decalin	di-naphthenes (C_nH_{2n-2})	99	Fisher Scientific
bicyclohexyl	di-naphthenes (C_nH_{2n-2})	99	Fisher Scientific
1,5-hexadiene	di-olefins (C_nH_{2n-2})	98	Fisher Scientific
1,7-octadiene	di-olefins (C_nH_{2n-2})	98.5	Fisher Scientific
2,4-dimethyl-1,3-pentadiene	di-olefins (C_nH_{2n-2})	98	Merck
cyclopentene	naphtheno-mono-olefins (C_nH_{2n-2})	>98	Fisher Scientific
cyclohexene	naphtheno-mono-olefins (C_nH_{2n-2})	99	Fisher Scientific
allylcyclohexane	naphtheno-mono-olefins (C_nH_{2n-2})	96	Merck
benzene	mono-aromatics (C_nH_{2n-6})	99.8	Fisher Scientific
toluene	mono-aromatics (C_nH_{2n-6})	99.85	Fisher Scientific
ethylbenzene	mono-aromatics (C_nH_{2n-6})	99.8	Fisher Scientific
m-xylene	mono-aromatics (C_nH_{2n-6})	>99	Fisher Scientific
p-xylene	mono-aromatics (C_nH_{2n-6})	99	Fisher Scientific
o-xylene	mono-aromatics (C_nH_{2n-6})	99	Fisher Scientific
mesitylene	mono-aromatics (C_nH_{2n-6})	99	Fisher Scientific
cumene	mono-aromatics (C_nH_{2n-6})	99.9	Fisher Scientific
n-propylbenzene	mono-aromatics (C_nH_{2n-6})	98	Fisher Scientific
1,2,4-trimethylbenzene	mono-aromatics (C_nH_{2n-6})	98	Fisher Scientific
sec-butylbenzene	mono-aromatics (C_nH_{2n-6})	99	Fisher Scientific
1,4-diethylbenzene	mono-aromatics (C_nH_{2n-6})	98	Fisher Scientific
n-butylbenzene	mono-aromatics (C_nH_{2n-6})	99	Fisher Scientific
p-cymene	mono-aromatics (C_nH_{2n-6})	>99	Fisher Scientific
isobutylbenzene	mono-aromatics (C_nH_{2n-6})	99	Merck
1,2,4,5-tetramethylbenzene	mono-aromatics (C_nH_{2n-6})	>97	Fisher Scientific
tert-butylbenzene	mono-aromatics (C_nH_{2n-6})	99	Fisher Scientific
n-pentylbenzene	mono-aromatics (C_nH_{2n-6})	96	Fisher Scientific
tert-pentylbenzene	mono-aromatics (C_nH_{2n-6})	97	Fisher Scientific
5-*tert*-butyl-*m*-xylene	mono-aromatics (C_nH_{2n-6})	98	Fisher Scientific
1,4-diisopropylbenzene	mono-aromatics (C_nH_{2n-6})	99	Fisher Scientific
n-hexylbenzene	mono-aromatics (C_nH_{2n-6})	98	Fisher Scientific
n-heptylbenzene	mono-aromatics (C_nH_{2n-6})	98	Fisher Scientific
n-octylbenzene	mono-aromatics (C_nH_{2n-6})	99	Fisher Scientific
1,4-di-*tert*-butylbenzene	mono-aromatics (C_nH_{2n-6})	98	Fisher Scientific
1,3,5-triisopropylbenzene	mono-aromatics (C_nH_{2n-6})	95	Fisher Scientific
indane	naphtheno-mono-aromatics (C_nH_{2n-8})	95	Fisher Scientific
tetralin	naphtheno-mono-aromatics (C_nH_{2n-8})	>98	Fisher Scientific
1,5-dimethyltetralin	naphtheno-mono-aromatics (C_nH_{2n-8})	>90	Merck
cyclohexylbenzene	naphtheno-mono-aromatics (C_nH_{2n-8})	98	Merck
allylbenzene	mono-olefin-aromatic (C_nH_{2n-8})	98	Fisher Scientific
naphthalene	di-aromatics (C_nH_{2n-12})	>99	Fisher Scientific
1-methylnaphthalene	di-aromatics (C_nH_{2n-12})	96	Fisher Scientific

Continued on next page

Table A.1 – continued from previous page.

Compound	Family	Purity (%)	Provider
2-methylnaphthalene	di-aromatics (C_nH_{2n-12})	97	Fisher Scientific
2-ethylnaphthalene	di-aromatics (C_nH_{2n-12})	>99	Merck
diphenylmethane	naphtheno-di-aromatics (C_nH_{2n-14})	>99	Fisher Scientific
biphenyl	di-aromatics (C_nH_{2n-14})	99	Fisher Scientific
acenaphthene	naphtheno-di-aromatics (C_nH_{2n-14})	99	Fisher Scientific
fluorene	naphtheno-di-aromatics (C_nH_{2n-16})	>98	Fisher Scientific

Appendix B

Full list of measured Induction Period values

Table B.1 – Comprehensive list of compounds analyzed in this study, including their SMILES representation, hydrocarbon family, carbon number, analysis temperature, and corresponding IP value.

Compound	T (°C)	SMILES	Hydrocarbon family	Carbon number	IP (h)
n-pentane	120	CCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	5	102.7
n-pentane	130	CCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	5	45.7
n-pentane	140	CCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	5	20.3
n-hexane	120	CCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	6	40.8
n-hexane	140	CCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	6	7.5
n-heptane	140	CCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	7	5.5
n-heptane	160	CCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	7	1.3
n-octane	120	CCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	8	67.0
n-octane	130	CCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	8	22.1
n-octane	140	CCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	8	5.0
n-octane	160	CCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	8	1.0
n-nonane	120	CCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	9	21.9
n-nonane	130	CCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	9	10.4
n-nonane	140	CCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	9	3.8
n-decane	120	CCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	10	27.8
n-decane	140	CCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	10	3.0
n-decane	160	CCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	10	0.7
n-undecane	140	CCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	11	2.9
n-undecane	160	CCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	11	0.6
n-dodecane	140	CCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	12	2.7
n-dodecane	160	CCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	12	0.6
n-tridecane	140	CCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	13	3.7
n-tridecane	160	CCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	13	0.7
n-tetradecane	120	CCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	14	16.8
n-tetradecane	140	CCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	14	2.5
n-tetradecane	160	CCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	14	0.5
n-pentadecane	120	CCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	15	11.1
n-pentadecane	140	CCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	15	2.4
n-hexadecane	140	CCCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	16	2.5
n-hexadecane	160	CCCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	16	0.6
n-heptadecane	120	CCCCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	17	13.4

Continued on next page

Table B.1 – continued from previous page.

Compound	T (°C)	SMILES	Hydrocarbon family	Carbon number	IP (h)
n-heptadecane	140	CCCCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	17	2.2
n-heptadecane	160	CCCCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	17	0.5
n-octadecane	140	CCCCCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	18	2.4
n-octadecane	160	CCCCCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	18	0.5
n-nonadecane	140	CCCCCCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	19	2.4
n-nonadecane	160	CCCCCCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	19	0.5
n-eicosane	120	CCCCCCCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	20	12.4
n-eicosane	140	CCCCCCCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	20	2.2
n-eicosane	160	CCCCCCCCCCCCCCCCCCCC	*n*-paraffins (*n*-C_nH_{2n+2})	20	0.5
2,3-dimethylbutane	100	CC(C)C(C)C	iso-paraffins (*i*-C_nH_{2n+2})	6	5.7
2,3-dimethylbutane	120	CC(C)C(C)C	iso-paraffins (*i*-C_nH_{2n+2})	6	1.8
2,3-dimethylbutane	140	CC(C)C(C)C	iso-paraffins (*i*-C_nH_{2n+2})	6	0.7
2,2-dimethylbutane	100	CCC(C)(C)C	iso-paraffins (*i*-C_nH_{2n+2})	6	171.2
2,2-dimethylbutane	140	CCC(C)(C)C	iso-paraffins (*i*-C_nH_{2n+2})	6	25.6
3-methylpentane	120	CCC(C)CC	iso-paraffins (*i*-C_nH_{2n+2})	6	21.8
3-methylpentane	140	CCC(C)CC	iso-paraffins (*i*-C_nH_{2n+2})	6	4.5
2-methylpentane	120	CCCC(C)C	iso-paraffins (*i*-C_nH_{2n+2})	6	11.0
2-methylpentane	140	CCCC(C)C	iso-paraffins (*i*-C_nH_{2n+2})	6	3.2
2,3,4-trimethylpentane	100	CC(C)C(C)C(C)C	iso-paraffins (*i*-C_nH_{2n+2})	8	10.4
2,3,4-trimethylpentane	140	CC(C)C(C)C(C)C	iso-paraffins (*i*-C_nH_{2n+2})	8	0.8
2,2,4-trimethylpentane	120	CC(C)CC(C)(C)C	iso-paraffins (*i*-C_nH_{2n+2})	8	75.8
2,2,4-trimethylpentane	140	CC(C)CC(C)(C)C	iso-paraffins (*i*-C_nH_{2n+2})	8	27.4
2,2,4-trimethylpentane	160	CC(C)CC(C)(C)C	iso-paraffins (*i*-C_nH_{2n+2})	8	6.3
2,5-dimethylhexane	100	CC(C)CCC(C)C	iso-paraffins (*i*-C_nH_{2n+2})	8	9.3
2,5-dimethylhexane	120	CC(C)CCC(C)C	iso-paraffins (*i*-C_nH_{2n+2})	8	2.0
2,5-dimethylhexane	140	CC(C)CCC(C)C	iso-paraffins (*i*-C_nH_{2n+2})	8	0.5
2-methylheptane	130	CCCCCC(C)C	iso-paraffins (*i*-C_nH_{2n+2})	8	3.1
2-methylheptane	140	CCCCCC(C)C	iso-paraffins (*i*-C_nH_{2n+2})	8	1.5
2-methylnonane	120	CCCCCCCC(C)C	iso-paraffins (*i*-C_nH_{2n+2})	10	7.7
2-methylnonane	140	CCCCCCCC(C)C	iso-paraffins (*i*-C_nH_{2n+2})	10	1.4
2-methylnonane	160	CCCCCCCC(C)C	iso-paraffins (*i*-C_nH_{2n+2})	10	0.4
2,2,4,6,6-pentamethylheptane	140	CC(CC(C)(C)C)CC(C)(C)C	iso-paraffins (*i*-C_nH_{2n+2})	12	4.1
2,2,4,6,6-pentamethylheptane	160	CC(CC(C)(C)C)CC(C)(C)C	iso-paraffins (*i*-C_nH_{2n+2})	12	1.0
isododecane	120	CCCCCCCCCC(C)C	iso-paraffins (*i*-C_nH_{2n+2})	12	17.1
isododecane	140	CCCCCCCCCC(C)C	iso-paraffins (*i*-C_nH_{2n+2})	12	3.9
2,2,4,4,6,8,8-heptamethylnonane	120	CC(CC(C)(C)C)CC(C)(C)CC(C)(C)C	iso-paraffins (*i*-C_nH_{2n+2})	16	6.9
2,2,4,4,6,8,8-heptamethylnonane	140	CC(CC(C)(C)C)CC(C)(C)CC(C)(C)C	iso-paraffins (*i*-C_nH_{2n+2})	16	1.4
2,2,4,4,6,8,8-heptamethylnonane	160	CC(CC(C)(C)C)CC(C)(C)CC(C)(C)C	iso-paraffins (*i*-C_nH_{2n+2})	16	0.4
cyclopentane	110	C1CCCC1	mono-naphthenes (C_nH_{2n})	5	57.6
cyclopentane	120	C1CCCC1	mono-naphthenes (C_nH_{2n})	5	27.1
cyclopentane	130	C1CCCC1	mono-naphthenes (C_nH_{2n})	5	8.6
cyclohexane	130	C1CCCCC1	mono-naphthenes (C_nH_{2n})	6	14.6
cyclohexane	140	C1CCCCC1	mono-naphthenes (C_nH_{2n})	6	6.8
cyclohexane	160	C1CCCCC1	mono-naphthenes (C_nH_{2n})	6	1.4
cycloheptane	140	C1CCCCCC1	mono-naphthenes (C_nH_{2n})	7	2.9
cycloheptane	160	C1CCCCCC1	mono-naphthenes (C_nH_{2n})	7	0.6
methylcyclohexane	120	CC1CCCCC1	mono-naphthenes (C_nH_{2n})	7	10.5
methylcyclohexane	140	CC1CCCCC1	mono-naphthenes (C_nH_{2n})	7	2.1
methylcyclohexane	160	CC1CCCCC1	mono-naphthenes (C_nH_{2n})	7	0.6
cyclooctane	120	C1CCCCCCC1	mono-naphthenes (C_nH_{2n})	8	6.7
cyclooctane	140	C1CCCCCCC1	mono-naphthenes (C_nH_{2n})	8	1.2
1,3-dimethylcyclohexane	140	CC1CCCC(C)C1	mono-naphthenes (C_nH_{2n})	8	1.8
1,3-dimethylcyclohexane	160	CC1CCCC(C)C1	mono-naphthenes (C_nH_{2n})	8	0.4
ethylcyclohexane	140	CCC1CCCCC1	mono-naphthenes (C_nH_{2n})	8	2.7
ethylcyclohexane	160	CCC1CCCCC1	mono-naphthenes (C_nH_{2n})	8	0.6

Continued on next page

TABLE B.1 – continued from previous page.

Compound	T (°C)	SMILES	Hydrocarbon family	Carbon number	IP (h)
1,2,4-trimethylcyclohexane	100	CC1CCC(C(C1)C)C	mono-naphthenes (C_nH_{2n})	9	4.6
1,2,4-trimethylcyclohexane	120	CC1CCC(C(C1)C)C	mono-naphthenes (C_nH_{2n})	9	1.6
1,2,4-trimethylcyclohexane	140	CC1CCC(C(C1)C)C	mono-naphthenes (C_nH_{2n})	9	0.4
n-propylcyclohexane	120	CCCC1CCCCC1	mono-naphthenes (C_nH_{2n})	9	9.7
n-propylcyclohexane	140	CCCC1CCCCC1	mono-naphthenes (C_nH_{2n})	9	1.6
n-propylcyclohexane	160	CCCC1CCCCC1	mono-naphthenes (C_nH_{2n})	9	0.3
tert-butylcyclohexane	140	CC(C)(C)C1CCCCC1	mono-naphthenes (C_nH_{2n})	10	6.6
tert-butylcyclohexane	160	CC(C)(C)C1CCCCC1	mono-naphthenes (C_nH_{2n})	10	1.4
n-butylcyclohexane	140	CCCCC1CCCCC1	mono-naphthenes (C_nH_{2n})	10	2.2
n-butylcyclohexane	160	CCCCC1CCCCC1	mono-naphthenes (C_nH_{2n})	10	0.4
1-hexene	100	C=CCCCC	mono-olefins (C_nH_{2n})	6	8.6
1-hexene	120	C=CCCCC	mono-olefins (C_nH_{2n})	6	1.6
1-hexene	140	C=CCCCC	mono-olefins (C_nH_{2n})	6	0.4
3,3-dimethyl-1-butene	100	CC(C)(C)C=C	mono-olefins (C_nH_{2n})	6	38.6
3,3-dimethyl-1-butene	120	CC(C)(C)C=C	mono-olefins (C_nH_{2n})	6	6.5
3,3-dimethyl-1-butene	130	CC(C)(C)C=C	mono-olefins (C_nH_{2n})	6	2.6
2,3-dimethyl-2-butene	40	CC(C)=C(C)C	mono-olefins (C_nH_{2n})	6	1.9
2,3-dimethyl-2-butene	60	CC(C)=C(C)C	mono-olefins (C_nH_{2n})	6	1.4
2,3-dimethyl-2-butene	80	CC(C)=C(C)C	mono-olefins (C_nH_{2n})	6	0.3
2,3-dimethyl-1-butene	80	CC(C)C(=C)C	mono-olefins (C_nH_{2n})	6	2.9
2,3-dimethyl-1-butene	100	CC(C)C(=C)C	mono-olefins (C_nH_{2n})	6	0.7
trans-2-hexene	80	CCC\C=C\C	mono-olefins (C_nH_{2n})	6	4.8
trans-2-hexene	100	CCC\C=C\C	mono-olefins (C_nH_{2n})	6	1.1
1-octene	100	C=CCCCCCC	mono-olefins (C_nH_{2n})	8	9.5
1-octene	120	C=CCCCCCC	mono-olefins (C_nH_{2n})	8	1.6
2,4,4-trimethyl-1-pentene	80	CC(=C)CC(C)(C)C	mono-olefins (C_nH_{2n})	8	21.0
2,4,4-trimethyl-1-pentene	100	CC(=C)CC(C)(C)C	mono-olefins (C_nH_{2n})	8	3.3
2,4,4-trimethyl-2-pentene	80	CC(=CC(C)(C)C)C	mono-olefins (C_nH_{2n})	8	6.0
2,4,4-trimethyl-2-pentene	100	CC(=CC(C)(C)C)C	mono-olefins (C_nH_{2n})	8	1.3
1-dodecene	100	C=CCCCCCCCCCC	mono-olefins (C_nH_{2n})	12	9.8
1-dodecene	120	C=CCCCCCCCCCC	mono-olefins (C_nH_{2n})	12	1.7
1-dodecene	140	C=CCCCCCCCCCC	mono-olefins (C_nH_{2n})	12	0.4
1-hexadecene	100	C=CCCCCCCCCCCCCCC	mono-olefins (C_nH_{2n})	16	8.4
1-hexadecene	120	C=CCCCCCCCCCCCCCC	mono-olefins (C_nH_{2n})	16	1.7
1-hexadecene	140	C=CCCCCCCCCCCCCCC	mono-olefins (C_nH_{2n})	16	0.4
cyclopentene	60	C1=CCCC1	naphtheno-mono-olefins (C_nH_{2n-2})	5	4.0
cyclopentene	80	C1=CCCC1	naphtheno-mono-olefins (C_nH_{2n-2})	5	1.0
cyclopentene	100	C1=CCCC1	naphtheno-mono-olefins (C_nH_{2n-2})	5	0.3
cyclohexene	100	C1CCC=CC1	naphtheno-mono-olefins (C_nH_{2n-2})	6	1.3
cyclohexene	120	C1CCC=CC1	naphtheno-mono-olefins (C_nH_{2n-2})	6	0.5
allylcyclohexane	100	C=CCC1CCCCC1	naphtheno-mono-olefins (C_nH_{2n-2})	9	2.5
allylcyclohexane	120	C=CCC1CCCCC1	naphtheno-mono-olefins (C_nH_{2n-2})	9	0.7
allylcyclohexane	140	C=CCC1CCCCC1	naphtheno-mono-olefins (C_nH_{2n-2})	9	0.2
decalin	120	C1CCC2CCCCC2C1	di-naphthenes (C_nH_{2n-2})	10	4.9
decalin	140	C1CCC2CCCCC2C1	di-naphthenes (C_nH_{2n-2})	10	1.0
decalin	160	C1CCC2CCCCC2C1	di-naphthenes (C_nH_{2n-2})	10	0.3
bicyclohexyl	120	C1CCC(CC1)C2CCCCC2	di-naphthenes (C_nH_{2n-2})	12	20.2
bicyclohexyl	140	C1CCC(CC1)C2CCCCC2	di-naphthenes (C_nH_{2n-2})	12	3.1
1,5-hexadiene	100	C=CCCC=C	di-olefins (C_nH_{2n-2})	6	13.5
1,5-hexadiene	120	C=CCCC=C	di-olefins (C_nH_{2n-2})	6	2.2
2,4-dimethyl-1,3-pentadiene	40	CC(=CC(=C)C)C	di-olefins (C_nH_{2n-2})	7	5.5
2,4-dimethyl-1,3-pentadiene	60	CC(=CC(=C)C)C	di-olefins (C_nH_{2n-2})	7	0.9
2,4-dimethyl-1,3-pentadiene	100	CC(=CC(=C)C)C	di-olefins (C_nH_{2n-2})	7	0.1
1,7-octadiene	100	C=CCCCCC=C	di-olefins (C_nH_{2n-2})	8	1.9
1,7-octadiene	120	C=CCCCCC=C	di-olefins (C_nH_{2n-2})	8	0.4

Continued on next page

Table B.1 – continued from previous page.

Compound	T (°C)	SMILES	Hydrocarbon family	Carbon number	IP (h)
benzene	140	C1=CC=CC=C1	mono-aromatics (C_nH_{2n-6})	6	20.3
benzene	160	C1=CC=CC=C1	mono-aromatics (C_nH_{2n-6})	6	15.3
toluene	120	Cc1ccccc1	mono-aromatics (C_nH_{2n-6})	7	26.7
toluene	140	Cc1ccccc1	mono-aromatics (C_nH_{2n-6})	7	19.0
toluene	160	Cc1ccccc1	mono-aromatics (C_nH_{2n-6})	7	14.5
o-xylene	120	Cc1c(C)cccc1	mono-aromatics (C_nH_{2n-6})	8	34.0
o-xylene	140	Cc1c(C)cccc1	mono-aromatics (C_nH_{2n-6})	8	10.8
o-xylene	160	Cc1c(C)cccc1	mono-aromatics (C_nH_{2n-6})	8	3.4
m-xylene	140	Cc1cc(C)ccc1	mono-aromatics (C_nH_{2n-6})	8	21.2
m-xylene	160	Cc1cc(C)ccc1	mono-aromatics (C_nH_{2n-6})	8	13.0
p-xylene	140	Cc1ccc(C)cc1	mono-aromatics (C_nH_{2n-6})	8	20.5
p-xylene	160	Cc1ccc(C)cc1	mono-aromatics (C_nH_{2n-6})	8	6.3
ethylbenzene	120	CCc1ccccc1	mono-aromatics (C_nH_{2n-6})	8	34.1
ethylbenzene	140	CCc1ccccc1	mono-aromatics (C_nH_{2n-6})	8	15.6
cumene	100	CC(C)c1ccccc1	mono-aromatics (C_nH_{2n-6})	9	1.9
cumene	120	CC(C)c1ccccc1	mono-aromatics (C_nH_{2n-6})	9	0.6
cumene	140	CC(C)c1ccccc1	mono-aromatics (C_nH_{2n-6})	9	0.2
mesitylene	140	Cc1cc(C)cc(C)c1	mono-aromatics (C_nH_{2n-6})	9	23.8
mesitylene	160	Cc1cc(C)cc(C)c1	mono-aromatics (C_nH_{2n-6})	9	12.6
1,2,4-trimethylbenzene	140	Cc1ccc(C)c(C)c1	mono-aromatics (C_nH_{2n-6})	9	19.9
1,2,4-trimethylbenzene	160	Cc1ccc(C)c(C)c1	mono-aromatics (C_nH_{2n-6})	9	4.9
n-propylbenzene	120	CCCc1ccccc1	mono-aromatics (C_nH_{2n-6})	9	9.7
n-propylbenzene	140	CCCc1ccccc1	mono-aromatics (C_nH_{2n-6})	9	2.4
n-propylbenzene	160	CCCc1ccccc1	mono-aromatics (C_nH_{2n-6})	9	0.7
p-cymene	100	c1cc(ccc1C(C)C)C	mono-aromatics (C_nH_{2n-6})	10	5.7
p-cymene	120	c1cc(ccc1C(C)C)C	mono-aromatics (C_nH_{2n-6})	10	2.0
p-cymene	140	c1cc(ccc1C(C)C)C	mono-aromatics (C_nH_{2n-6})	10	0.4
tert-butylbenzene	140	CC(C)(C)c1ccccc1	mono-aromatics (C_nH_{2n-6})	10	29.6
tert-butylbenzene	160	CC(C)(C)c1ccccc1	mono-aromatics (C_nH_{2n-6})	10	19.9
isobutylbenzene	140	CC(C)Cc1ccccc1	mono-aromatics (C_nH_{2n-6})	10	4.3
isobutylbenzene	160	CC(C)Cc1ccccc1	mono-aromatics (C_nH_{2n-6})	10	0.9
1,2,4,5-tetramethylbenzene	140	Cc1cc(C)c(C)cc1C	mono-aromatics (C_nH_{2n-6})	10	1.8
sec-butylbenzene	120	CCC(C)c1ccccc1	mono-aromatics (C_nH_{2n-6})	10	8.9
sec-butylbenzene	140	CCC(C)c1ccccc1	mono-aromatics (C_nH_{2n-6})	10	2.7
1,4-diethylbenzene	120	CCC1=CC=C(C=C1)CC	mono-aromatics (C_nH_{2n-6})	10	4.5
1,4-diethylbenzene	140	CCC1=CC=C(C=C1)CC	mono-aromatics (C_nH_{2n-6})	10	2.9
n-butylbenzene	120	CCCCc1ccccc1	mono-aromatics (C_nH_{2n-6})	10	12.3
n-butylbenzene	140	CCCCc1ccccc1	mono-aromatics (C_nH_{2n-6})	10	4.0
n-butylbenzene	160	CCCCc1ccccc1	mono-aromatics (C_nH_{2n-6})	10	1.2
pentamethylbenzene	140	Cc1cc(C)c(C)c(C)c1C	mono-aromatics (C_nH_{2n-6})	11	0.8
tert-pentylbenzene	140	CCC(C)(C)C1=CC=CC=C1	mono-aromatics (C_nH_{2n-6})	11	31.2
tert-pentylbenzene	160	CCC(C)(C)C1=CC=CC=C1	mono-aromatics (C_nH_{2n-6})	11	19.6
1,4-diisopropylbenzene	80	CC(C)c1ccc(cc1)C(C)C	mono-aromatics (C_nH_{2n-6})	12	4.3
1,4-diisopropylbenzene	100	CC(C)c1ccc(cc1)C(C)C	mono-aromatics (C_nH_{2n-6})	12	1.1
1,4-diisopropylbenzene	120	CC(C)c1ccc(cc1)C(C)C	mono-aromatics (C_nH_{2n-6})	12	0.4
5-*tert*-butyl-*m*-xylene	140	Cc1cc(C)cc(c1)C(C)(C)C	mono-aromatics (C_nH_{2n-6})	12	23.5
5-*tert*-butyl-*m*-xylene	160	Cc1cc(C)cc(c1)C(C)(C)C	mono-aromatics (C_nH_{2n-6})	12	15.6
hexylbenzene	140	CCCCCCc1ccccc1	mono-aromatics (C_nH_{2n-6})	12	2.2
hexylbenzene	160	CCCCCCc1ccccc1	mono-aromatics (C_nH_{2n-6})	12	0.6
1-phenylheptane	120	CCCCCCCc1ccccc1	mono-aromatics (C_nH_{2n-6})	13	12.8
1-phenylheptane	140	CCCCCCCc1ccccc1	mono-aromatics (C_nH_{2n-6})	13	2.8
1,4-di-*tert*-butylbenzene	140	CC(C)(C)c1ccc(cc1)C(C)(C)C	mono-aromatics (C_nH_{2n-6})	14	39.7
n-octylbenzene	140	CCCCCCCCc1ccccc1	mono-aromatics (C_nH_{2n-6})	14	3.2
n-octylbenzene	160	CCCCCCCCc1ccccc1	mono-aromatics (C_nH_{2n-6})	14	0.8
1,3,5-triisopropylbenzene	80	CC(C)c1cc(cc(c1)C(C)C)C(C)C	mono-aromatics (C_nH_{2n-6})	15	20.6

Continued on next page

Table B.1 – continued from previous page.

Compound	T (°C)	SMILES	Hydrocarbon family	Carbon number	IP (h)
1,3,5-triisopropylbenzene	100	CC(C)c1cc(cc(c1)C(C)C)C(C)C	mono-aromatics (C_nH_{2n-6})	15	4.1
indane	100	C1Cc2ccccc2C1	naphtheno-mono-aromatics (C_nH_{2n-8})	9	3.1
indane	120	C1Cc2ccccc2C1	naphtheno-mono-aromatics (C_nH_{2n-8})	9	0.8
indane	140	C1Cc2ccccc2C1	naphtheno-mono-aromatics (C_nH_{2n-8})	9	0.2
tetralin	100	C1CCc2ccccc2C1	naphtheno-mono-aromatics (C_nH_{2n-8})	10	15.1
tetralin	120	C1CCc2ccccc2C1	naphtheno-mono-aromatics (C_nH_{2n-8})	10	3.4
tetralin	140	C1CCc2ccccc2C1	naphtheno-mono-aromatics (C_nH_{2n-8})	10	0.8
cyclohexylbenzene	120	C1CCC(CC1)c2ccccc2	naphtheno-mono-aromatics (C_nH_{2n-8})	12	5.9
cyclohexylbenzene	140	C1CCC(CC1)c2ccccc2	naphtheno-mono-aromatics (C_nH_{2n-8})	12	1.3
1,5-dimethyltetralin	80	CC1CCCc2c(C)cccc12	naphtheno-mono-aromatics (C_nH_{2n-8})	12	1.7
1,5-dimethyltetralin	100	CC1CCCc2c(C)cccc12	naphtheno-mono-aromatics (C_nH_{2n-8})	12	0.4
1,5-dimethyltetralin	120	CC1CCCc2c(C)cccc12	naphtheno-mono-aromatics (C_nH_{2n-8})	12	0.1
allylbenzene	80	C=CCC1=CC=CC=C1	mono-olefin-aromatic (C_nH_{2n-8})	9	1.9
allylbenzene	100	C=CCC1=CC=CC=C1	mono-olefin-aromatic (C_nH_{2n-8})	9	0.9
allylbenzene	120	C=CCC1=CC=CC=C1	mono-olefin-aromatic (C_nH_{2n-8})	9	0.3
naphthalene	140	c1ccc2ccccc2c1	di-aromatics (C_nH_{2n-12})	10	25.4
2-methylnaphthalene	140	Cc1ccc2ccccc2c1	di-aromatics (C_nH_{2n-12})	11	28.6
1-methylnaphthalene	140	Cc1cccc2ccccc12	di-aromatics (C_nH_{2n-12})	11	30.8
1-methylnaphthalene	160	Cc1cccc2ccccc12	di-aromatics (C_nH_{2n-12})	11	14.7
2-ethylnaphthalene	140	CCc1ccc2ccccc2c1	di-aromatics (C_nH_{2n-12})	12	22.9
2-ethylnaphthalene	160	CCc1ccc2ccccc2c1	di-aromatics (C_nH_{2n-12})	12	9.1
biphenyl	140	c1ccc(cc1)c2ccccc2	di-aromatics (C_nH_{2n-14})	12	33.5
diphenylmethane	120	C(c1ccccc1)c2ccccc2	di-aromatics (C_nH_{2n-14})	13	2.9
diphenylmethane	140	C(c1ccccc1)c2ccccc2	di-aromatics (C_nH_{2n-14})	13	1.2
acenaphthene	120	C1Cc2cccc3cccc1c23	naphtheno-di-aromatics (C_nH_{2n-14})	12	6.3
acenaphthene	140	C1Cc2cccc3cccc1c23	naphtheno-di-aromatics (C_nH_{2n-14})	12	1.0
fluorene	140	C1c2ccccc2c3ccccc13	naphtheno-di-aromatics (C_nH_{2n-16})	13	0.5

Appendix C

MaxMin Maximum-Dissimilarity Algorithm

The MaxMin Maximum-Dissimilarity Algorithm was first described by Kennard and Stone[237]. Consider the problem of uniformly selecting a subset of k samples from a set of n samples. The matrix representing the dataset is given by:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{C.1}$$

Where:

- n is the number of samples and
- m the number of variables.

The first step of the algorithm consists of calculating a distance matrix **D**:

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} \tag{C.2}$$

Containing the distance of each pair of samples in the datasets, regularly, the Euclidean distance is used.

The second step is the algorithm's initialization by choosing an initial sample. The selection can be made in several ways: Random selection, selecting the most dissimilar compound with respect to the others, or choosing the point closest to the dataset center [312]. In the third

step, the compound with the largest dissimilarity with respect to the initial point is added to the subset. Then, in the fourth step, the maximum minimum dissimilarity criterion is used; for a given sample i in the subset, the distance between it and every other sample in the set is calculated, then the minimum distance is kept. The process is repeated for every point in the subset. Afterward, the sample with the maximum minimum distance will be selected and added to the subset. The process is repeated until the number of compounds in the subset = k.

Appendix D

List of molecules used for Data Augmentation

Table D.1 – List of compounds of molecules used for Data Augmentation, including their SMILES representation, hydrocarbon family, carbon number, analysis temperature, and corresponding IP value.

Compound	T (°C)	SMILES	Hydrocarbon family	Carbon number	IP (h)
1-*tert*-butyl-3-methylcyclohexane	120	CC1CCCC(C1)C(C)(C)C	mono-naphthenes (C_nH_{2n})	11	10.5
1-*tert*-butyl-3-methylcyclohexane	140	CC1CCCC(C1)C(C)(C)C	mono-naphthenes (C_nH_{2n})	11	2.1
1-*tert*-butyl-3-methylcyclohexane	160	CC1CCCC(C1)C(C)(C)C	mono-naphthenes (C_nH_{2n})	11	0.6
1-*tert*-butyl-4-methylcyclohexane	120	CC1CCC(CC1)C(C)(C)C	mono-naphthenes (C_nH_{2n})	11	10.5
1-*tert*-butyl-4-methylcyclohexane	140	CC1CCC(CC1)C(C)(C)C	mono-naphthenes (C_nH_{2n})	11	2.1
1-*tert*-butyl-4-methylcyclohexane	160	CC1CCC(CC1)C(C)(C)C	mono-naphthenes (C_nH_{2n})	11	0.6
6-methyl-1,2,3,4-tetrahydronaphthalene	100	c1c(C)cc2c(c1)CCCC2	naphtheno-mono-aromatics (C_nH_{2n-8})	11	15.1
6-methyl-1,2,3,4-tetrahydronaphthalene	120	c1c(C)cc2c(c1)CCCC2	naphtheno-mono-aromatics (C_nH_{2n-8})	11	3.4
6-methyl-1,2,3,4-tetrahydronaphthalene	140	c1c(C)cc2c(c1)CCCC2	naphtheno-mono-aromatics (C_nH_{2n-8})	11	0.8
5-methylindan	100	c1c(C)cc2c(c1)CCC2	naphtheno-mono-aromatics (C_nH_{2n-8})	10	3.1
5-methylindan	120	c1c(C)cc2c(c1)CCC2	naphtheno-mono-aromatics (C_nH_{2n-8})	10	0.8
5-methylindan	140	c1c(C)cc2c(c1)CCC2	naphtheno-mono-aromatics (C_nH_{2n-8})	10	0.2
2-methylfluorene	140	C1c2cc(C)ccc2c3ccccc13	naphtheno-di-aromatics (C_nH_{2n-16})	14	0.5
3-methylfluorene	140	CC1=CC2=C(CC3=CC=CC=C32)C=C1	naphtheno-di-aromatics (C_nH_{2n-16})	14	0.5
4-methylacenaphthene	120	CC1=CC2=C3C(=CC=C2)CCC3=C1	naphtheno-di-aromatics (C_nH_{2n-14})	13	6.3
4-methylacenaphthene	140	CC1=CC2=C3C(=CC=C2)CCC3=C1	naphtheno-di-aromatics (C_nH_{2n-14})	13	1.0
5-methylacenaphthene	120	CC1=CC=C2CCC3=C2C1=CC=C3	naphtheno-di-aromatics (C_nH_{2n-14})	13	6.3
5-methylacenaphthene	140	CC1=CC=C2CCC3=C2C1=CC=C3	naphtheno-di-aromatics (C_nH_{2n-14})	13	1.0
3-methyldiphenylmethane	120	CC1=CC(=CC=C1)CC2=CC=CC=C2	di-aromatics (C_nH_{2n-14})	14	2.9
3-methyldiphenylmethane	140	CC1=CC(=CC=C1)CC2=CC=CC=C2	di-aromatics (C_nH_{2n-14})	14	1.2
4-methyldiphenylmethane	120	CC1=CC=C(C=C1)CC2=CC=CC=C2	di-aromatics (C_nH_{2n-14})	14	2.9
4-methyldiphenylmethane	140	CC1=CC=C(C=C1)CC2=CC=CC=C2	di-aromatics (C_nH_{2n-14})	14	1.2
2-methylhexane	140	CCCCC(C)C	iso-paraffins (i-C_nH_{2n+2})	7	2.5
2-methylhexane	160	CCCCC(C)C	iso-paraffins (i-C_nH_{2n+2})	7	0.6
2-methyloctane	120	CCCCCCC(C)C	iso-paraffins (i-C_nH_{2n+2})	9	9.8

Continued on next page

TABLE D.1 – continued from previous page.

Compound	T (°C)	SMILES	Hydrocarbon family	Carbon number	IP (h)
2-methyloctane	130	CCCCCCC(C)C	iso-paraffins (i-C_nH_{2n+2})	9	4.7
2-methyloctane	140	CCCCCCC(C)C	iso-paraffins (i-C_nH_{2n+2})	9	1.7
2-methyldecane	140	CCCCCCCCC(C)C	iso-paraffins (i-C_nH_{2n+2})	11	1.3
2-methyldecane	160	CCCCCCCCC(C)C	iso-paraffins (i-C_nH_{2n+2})	11	0.3
2,4,4-trimethyl-1-hexene	80	CCC(C)(C)CC(=C)C	mono-olefins (C_nH_{2n})	9	21.0
2,4,4-trimethyl-1-hexene	100	CCC(C)(C)CC(=C)C	mono-olefins (C_nH_{2n})	9	3.3
2,2,4,6,6,8,8-heptamethyldecane	120	CC(CC(C)(C)C)CC(C)(C)CC(C)(C)CC	iso-paraffins (i-C_nH_{2n+2})	17	6.9
2,2,4,6,6,8,8-heptamethyldecane	140	CC(CC(C)(C)C)CC(C)(C)CC(C)(C)CC	iso-paraffins (i-C_nH_{2n+2})	17	1.4
2,2,4,6,6,8,8-heptamethyldecane	160	CC(CC(C)(C)C)CC(C)(C)CC(C)(C)CC	iso-paraffins (i-C_nH_{2n+2})	17	0.4
2,2,4,4,6,8,8-heptamethyldecane	120	CC(CC(CC)(C)C)CC(C)(C)CC(C)(C)C	iso-paraffins (i-C_nH_{2n+2})	17	6.9
2,2,4,4,6,8,8-heptamethyldecane	140	CC(CC(CC)(C)C)CC(C)(C)CC(C)(C)C	iso-paraffins (i-C_nH_{2n+2})	17	1.4
2,2,4,4,6,8,8-heptamethyldecane	160	CC(CC(CC)(C)C)CC(C)(C)CC(C)(C)C	iso-paraffins (i-C_nH_{2n+2})	17	0.4
4-allyltoluene	80	CC1=CC=C(C=C1)CC=C	mono-olefin-aromatic (C_nH_{2n-8})	10	1.9
4-allyltoluene	100	CC1=CC=C(C=C1)CC=C	mono-olefin-aromatic (C_nH_{2n-8})	10	0.9
4-allyltoluene	120	CC1=CC=C(C=C1)CC=C	mono-olefin-aromatic (C_nH_{2n-8})	10	0.3
m-allyltoluene	80	CC1=CC(=CC=C1)CC=C	mono-olefin-aromatic (C_nH_{2n-8})	10	1.9
m-allyltoluene	100	CC1=CC(=CC=C1)CC=C	mono-olefin-aromatic (C_nH_{2n-8})	10	0.9
m-allyltoluene	120	CC1=CC(=CC=C1)CC=C	mono-olefin-aromatic (C_nH_{2n-8})	10	0.3
2,2,4-trimethylhexane	120	CCC(C)CC(C)(C)C	iso-paraffins (i-C_nH_{2n+2})	9	68.3
2,2,4-trimethylhexane	140	CCC(C)CC(C)(C)C	iso-paraffins (i-C_nH_{2n+2})	9	24.6
2,2,4-trimethylhexane	160	CCC(C)CC(C)(C)C	iso-paraffins (i-C_nH_{2n+2})	9	5.7
2,4,4-trimethylhexane	120	CC(C)CC(C)(C)CC	iso-paraffins (i-C_nH_{2n+2})	9	68.3
2,4,4-trimethylhexane	140	CC(C)CC(C)(C)CC	iso-paraffins (i-C_nH_{2n+2})	9	24.6
2,4,4-trimethylhexane	160	CC(C)CC(C)(C)CC	iso-paraffins (i-C_nH_{2n+2})	9	5.7
2,2,4,6,6-pentamethyloctane	140	CCC(C)(C)CC(C)CC(C)(C)C	iso-paraffins (i-C_nH_{2n+2})	13	3.7
2,2,4,6,6-pentamethyloctane	160	CCC(C)(C)CC(C)CC(C)(C)C	iso-paraffins (i-C_nH_{2n+2})	13	0.9
4-ethyl-2,2,6,6-tetramethylheptane	140	CCC(CC(C)(C)C)CC(C)(C)C	iso-paraffins (i-C_nH_{2n+2})	13	3.7
4-ethyl-2,2,6,6-tetramethylheptane	160	CCC(CC(C)(C)C)CC(C)(C)C	iso-paraffins (i-C_nH_{2n+2})	13	0.9
3,3-dimethyl-1-pentene	100	CCC(C)(C)C=C	mono-olefins (C_nH_{2n})	7	34.7
3,3-dimethyl-1-pentene	120	CCC(C)(C)C=C	mono-olefins (C_nH_{2n})	7	5.9
3,3-dimethyl-1-pentene	130	CCC(C)(C)C=C	mono-olefins (C_nH_{2n})	7	2.3
1-sec-butyl-3-methylbenzene	120	CCC(C)C1=CC=CC(=C1)C	mono-aromatics (C_nH_{2n-6})	11	8.9
1-sec-butyl-3-methylbenzene	140	CCC(C)C1=CC=CC(=C1)C	mono-aromatics (C_nH_{2n-6})	11	2.7
1-sec-butyl-4-methylbenzene	120	CCC(C)C1=CC=C(C=C1)C	mono-aromatics (C_nH_{2n-6})	11	8.9
1-sec-butyl-4-methylbenzene	140	CCC(C)C1=CC=C(C=C1)C	mono-aromatics (C_nH_{2n-6})	11	2.7
1,4-diethyl-2-methylbenzene	120	CCC1=CC(=C(C=C1)CC)C	mono-aromatics (C_nH_{2n-6})	11	4.5
1,4-diethyl-2-methylbenzene	140	CCC1=CC(=C(C=C1)CC)C	mono-aromatics (C_nH_{2n-6})	11	2.9
1-butyl-3-methylbenzene	120	CCCCC1=CC=CC(=C1)C	mono-aromatics (C_nH_{2n-6})	11	12.3
1-butyl-3-methylbenzene	140	CCCCC1=CC=CC(=C1)C	mono-aromatics (C_nH_{2n-6})	11	4.0
1-butyl-3-methylbenzene	160	CCCCC1=CC=CC(=C1)C	mono-aromatics (C_nH_{2n-6})	11	1.2
1-butyl-4-methylbenzene	120	CCCCC1=CC=C(C=C1)C	mono-aromatics (C_nH_{2n-6})	11	12.3
1-butyl-4-methylbenzene	140	CCCCC1=CC=C(C=C1)C	mono-aromatics (C_nH_{2n-6})	11	4.0
1-butyl-4-methylbenzene	160	CCCCC1=CC=C(C=C1)C	mono-aromatics (C_nH_{2n-6})	11	1.2
1,4-diisopropyl-2-methylbenzene	80	CC1=C(C=CC(=C1)C(C)C)C(C)C	mono-aromatics (C_nH_{2n-6})	13	4.3
1,4-diisopropyl-2-methylbenzene	100	CC1=C(C=CC(=C1)C(C)C)C(C)C	mono-aromatics (C_nH_{2n-6})	13	1.1
1,4-diisopropyl-2-methylbenzene	120	CC1=C(C=CC(=C1)C(C)C)C(C)C	mono-aromatics (C_nH_{2n-6})	13	0.4
1,3,5-triisopropyl-2-methylbenzene	80	CC1=C(C=C(C=C1C(C)C)C(C)C)C(C)C	mono-aromatics (C_nH_{2n-6})	16	20.6
1,3,5-triisopropyl-2-methylbenzene	100	CC1=C(C=C(C=C1C(C)C)C(C)C)C(C)C	mono-aromatics (C_nH_{2n-6})	16	4.1
1-cyclohexyl-3-methylbenzene	120	CC1=CC(=CC=C1)C2CCCCC2	naphtheno-mono-aromatics (C_nH_{2n-8})	13	5.9
1-cyclohexyl-3-methylbenzene	140	CC1=CC(=CC=C1)C2CCCCC2	naphtheno-mono-aromatics (C_nH_{2n-8})	13	1.3
1-cyclohexyl-4-methylbenzene	120	CC1=CC=C(C=C1)C2CCCCC2	naphtheno-mono-aromatics (C_nH_{2n-8})	13	5.9

Continued on next page

Table D.1 – continued from previous page.

Compound	T (°C)	SMILES	Hydrocarbon family	Carbon number	IP (h)
1-cyclohexyl-4-methylbenzene	140	CC1=CC=C(C=C1)C2CCCCC2	naphtheno-mono-aromatics (C_nH_{2n-8})	13	1.3
1-methyl-4-propylbenzene	120	CCCC1=CC=C(C=C1)C	mono-aromatics (C_nH_{2n-6})	10	9.7
1-methyl-4-propylbenzene	140	CCCC1=CC=C(C=C1)C	mono-aromatics (C_nH_{2n-6})	10	2.4
1-methyl-4-propylbenzene	160	CCCC1=CC=C(C=C1)C	mono-aromatics (C_nH_{2n-6})	10	0.7
1-methyl-3-propylbenzene	120	CCCC1=CC=CC(=C1)C	mono-aromatics (C_nH_{2n-6})	10	9.7
1-methyl-3-propylbenzene	140	CCCC1=CC=CC(=C1)C	mono-aromatics (C_nH_{2n-6})	10	2.4
1-methyl-3-propylbenzene	160	CCCC1=CC=CC(=C1)C	mono-aromatics (C_nH_{2n-6})	10	0.7
2,2-dimethylpentane	140	CCCC(C)(C)C	iso-paraffins (i-C_nH_{2n+2})	7	23.1

Appendix E

PCA scores and loadings plots for NIR spectroscopic data

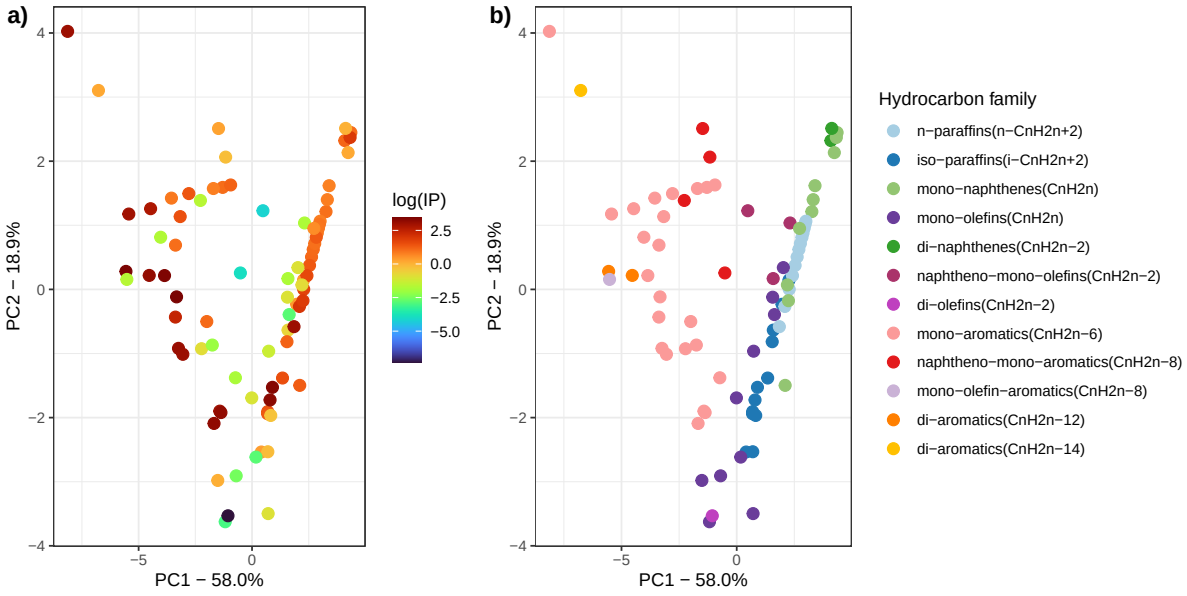


Figure E.1 – Scores plot for PC1 vs. PC2 based on raw NIR spectra. The data points are color-coded to represent **a)** log(IP) values and **b)** hydrocarbon families.

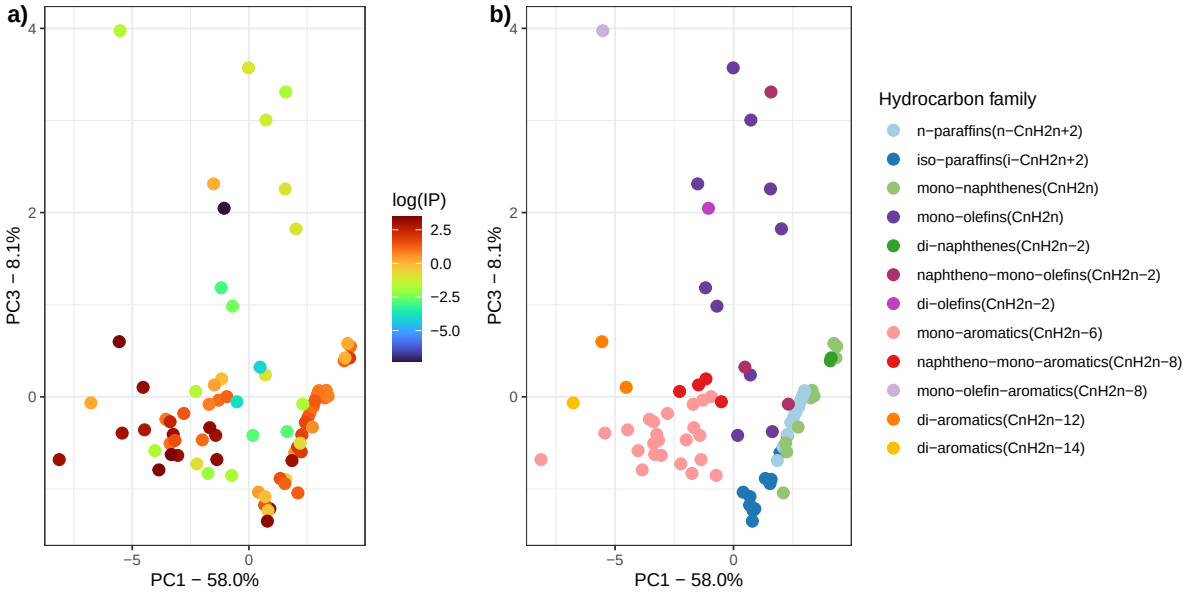


FIGURE E.2 – Scores plot for PC1 vs. PC3 based on raw NIR spectra. The data points are color-coded to represent **a)** log(IP) values and **b)** hydrocarbon families.

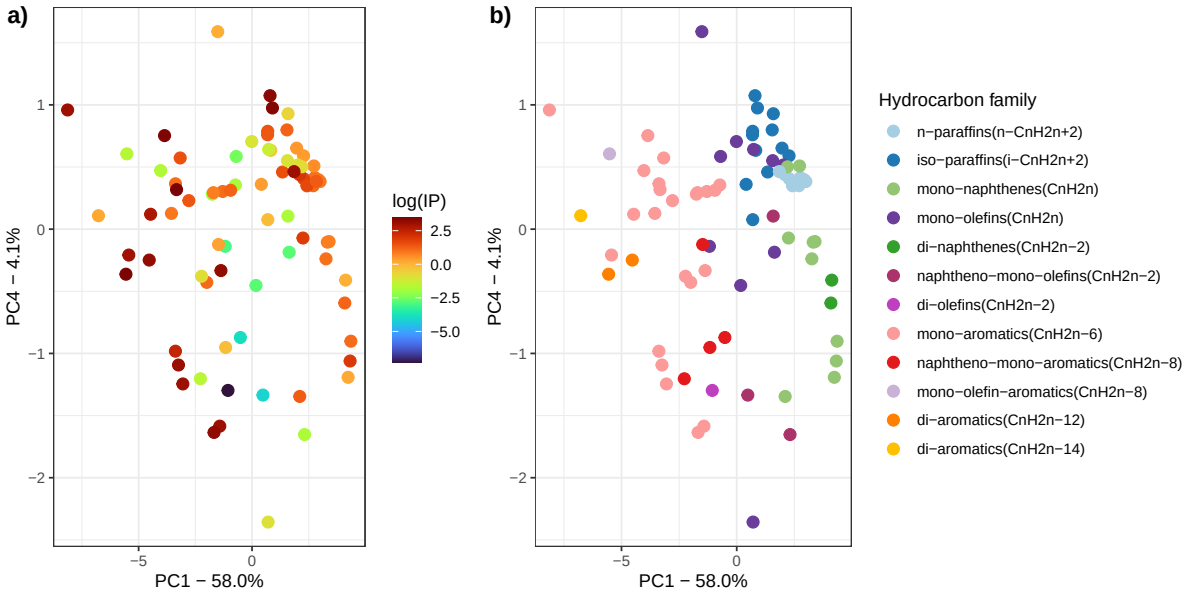


FIGURE E.3 – Scores plot for PC1 vs. PC4 based on raw NIR spectra. The data points are color-coded to represent **a)** log(IP) values and **b)** hydrocarbon families.

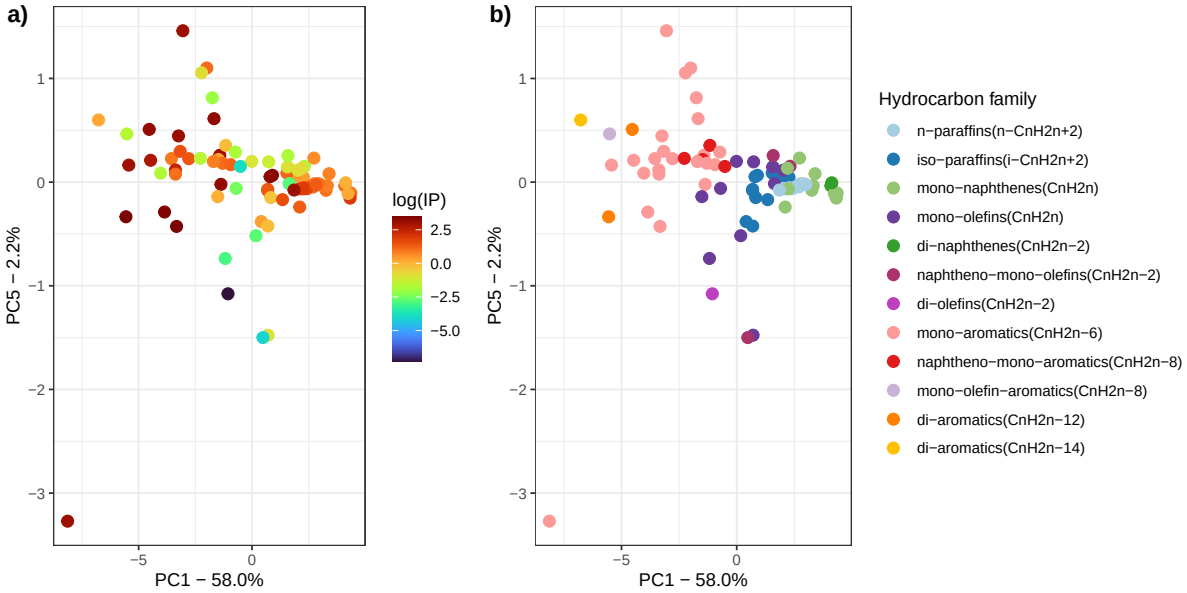


FIGURE E.4 – Scores plot for PC1 vs. PC5 based on raw NIR spectra. The data points are color-coded to represent **a)** log(IP) values and **b)** hydrocarbon families.

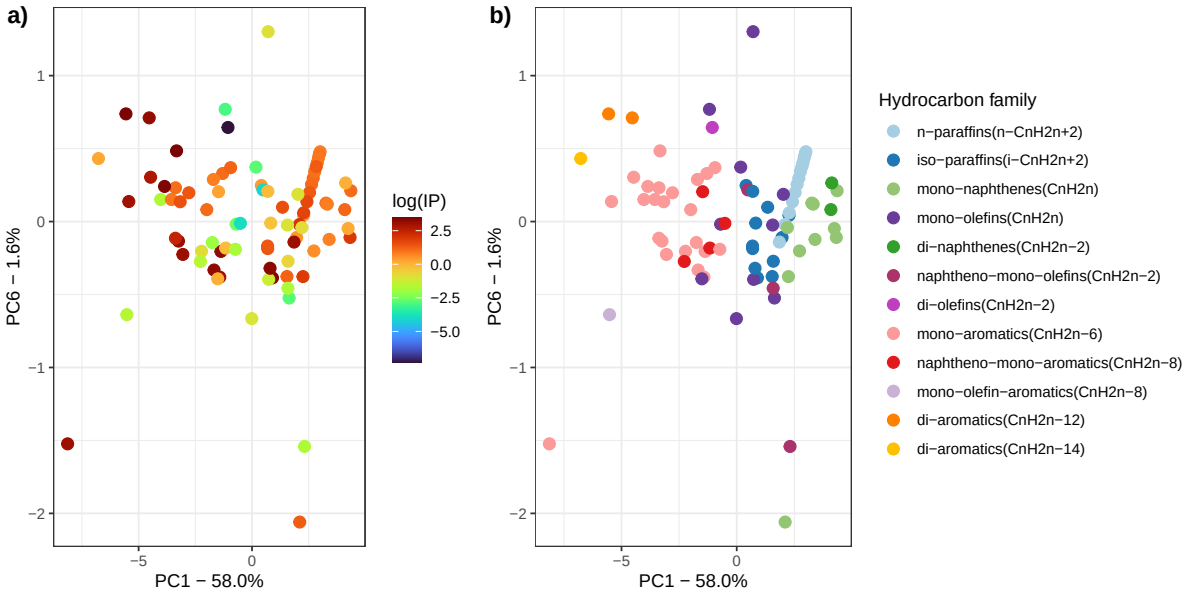


FIGURE E.5 – Scores plot for PC1 vs. PC6 based on raw NIR spectra. The data points are color-coded to represent **a)** log(IP) values and **b)** hydrocarbon families.

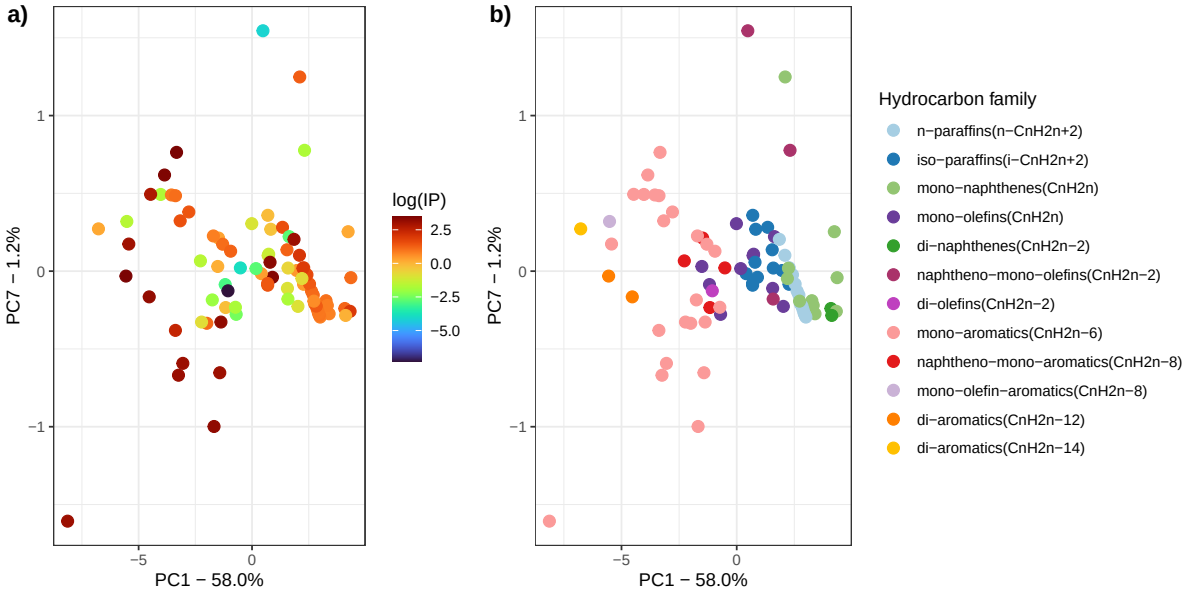


FIGURE E.6 – Scores plot for PC1 vs. PC7 based on raw NIR spectra. The data points are color-coded to represent **a)** log(IP) values and **b)** hydrocarbon families.

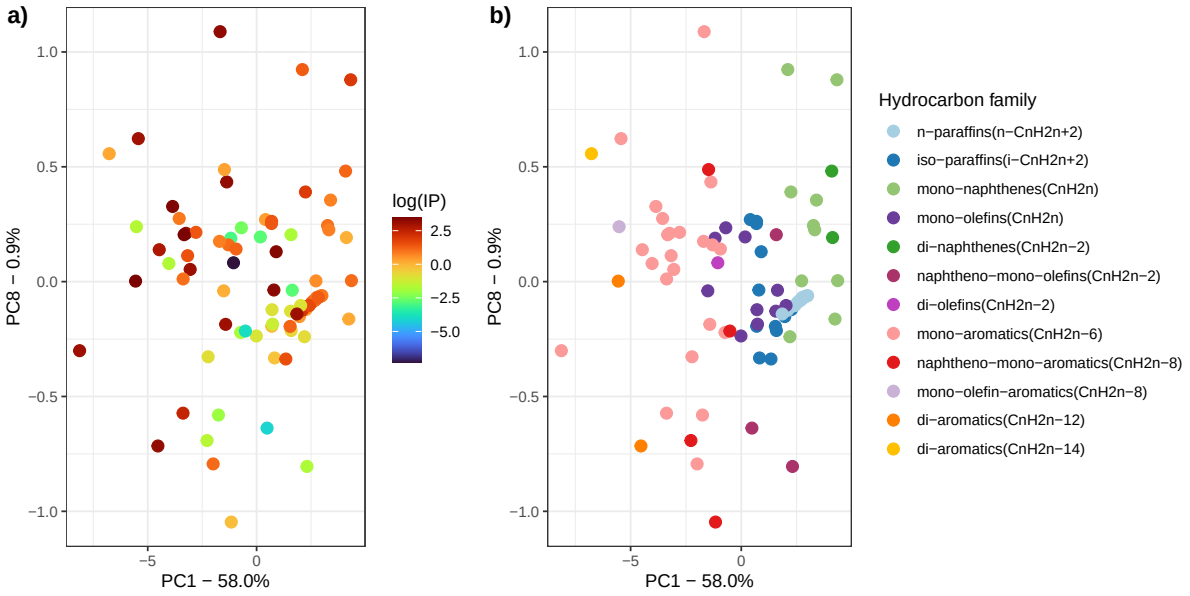


FIGURE E.7 – Scores plot for PC1 vs. PC8 based on raw NIR spectra. The data points are color-coded to represent **a)** log(IP) values and **b)** hydrocarbon families.

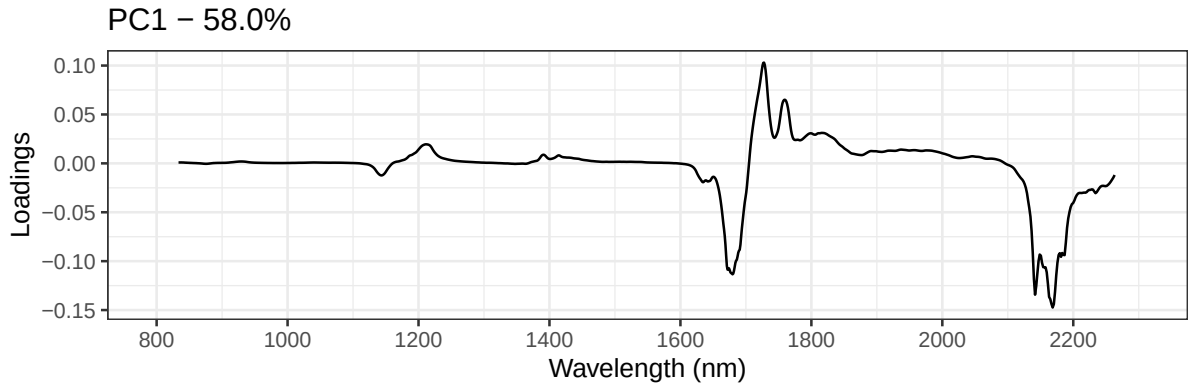


FIGURE E.8 – Loadings for PC1, which explains 58.0% of the variance.

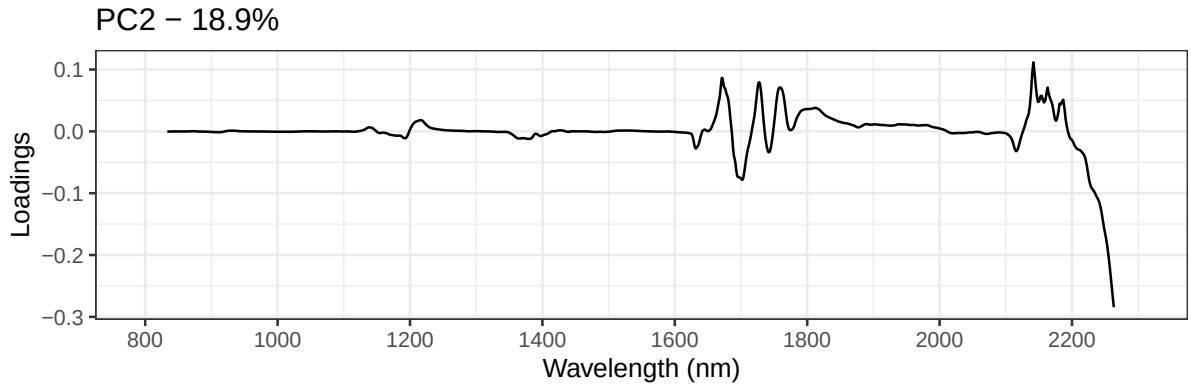


FIGURE E.9 – Loadings for PC2, which explains 18.9% of the variance.

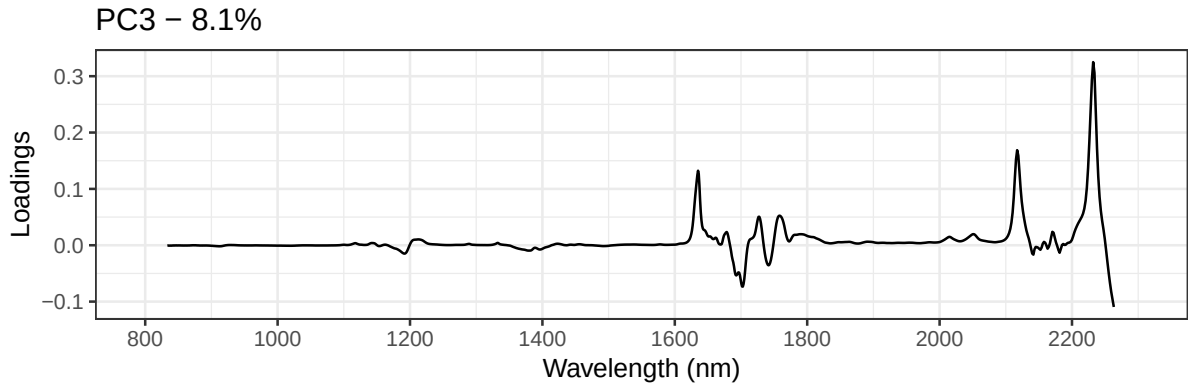


FIGURE E.10 – Loadings for PC3, which explains 8.1% of the variance.

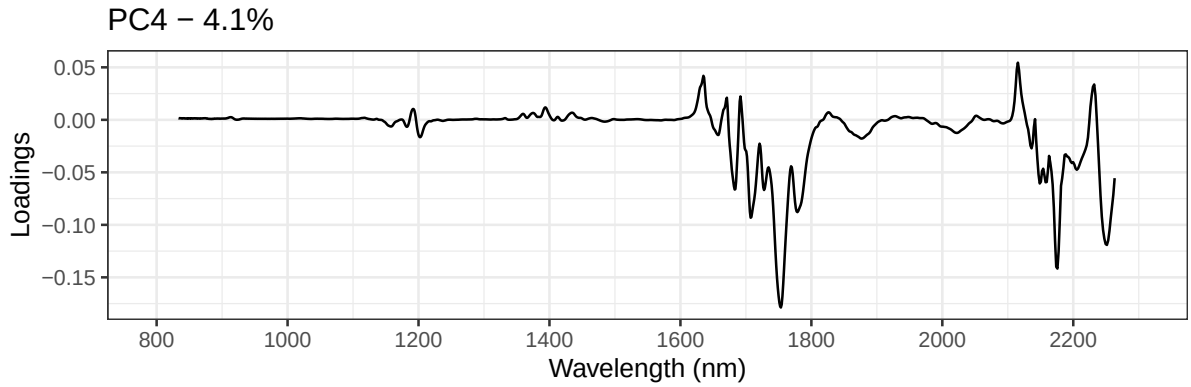


FIGURE E.11 – Loadings for PC4, which explains 4.1% of the variance.

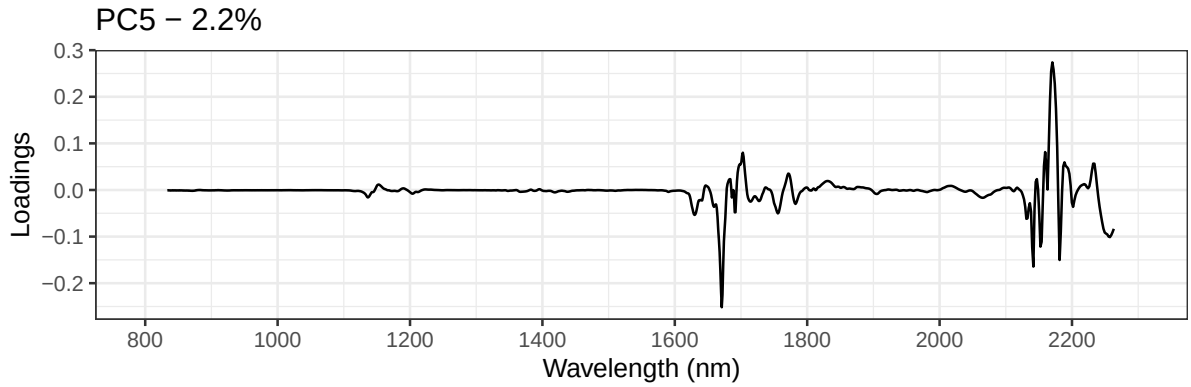


FIGURE E.12 – Loadings for PC5, which explains 2.2% of the variance.

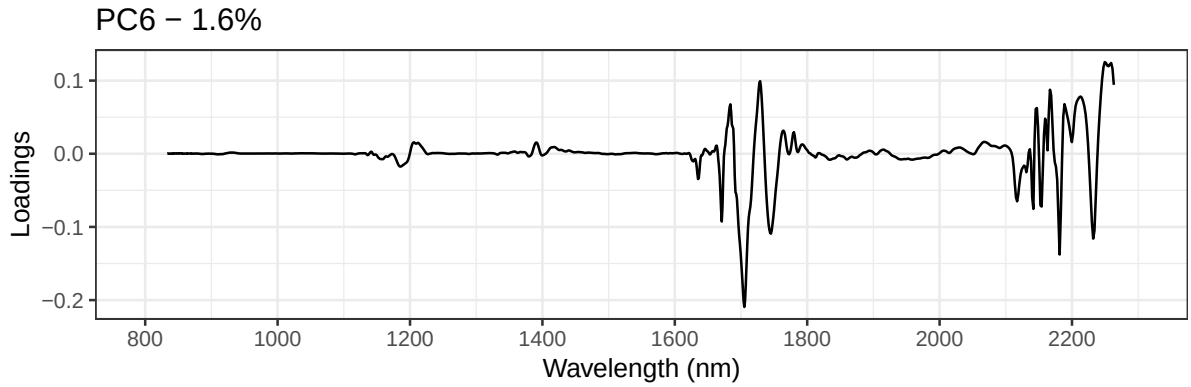


FIGURE E.13 – Loadings for PC6, which explains 1.6% of the variance.

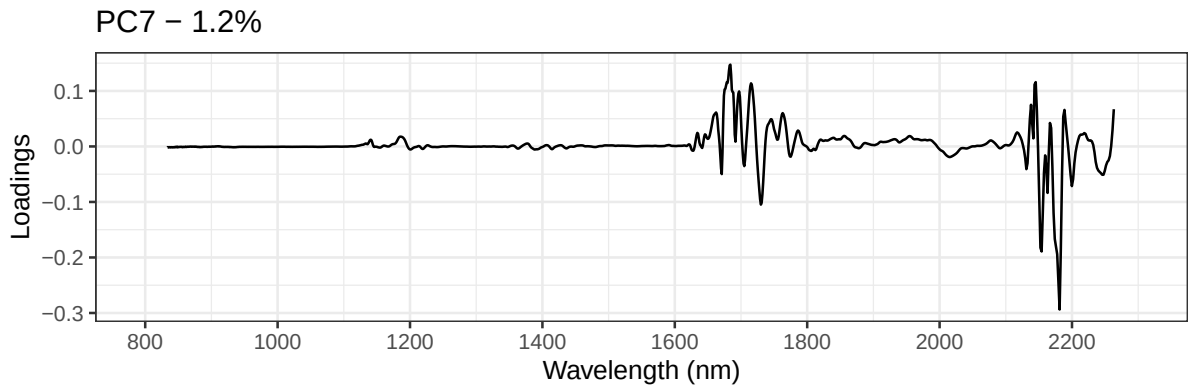


FIGURE E.14 – Loadings for PC7, which explains 1.2% of the variance.

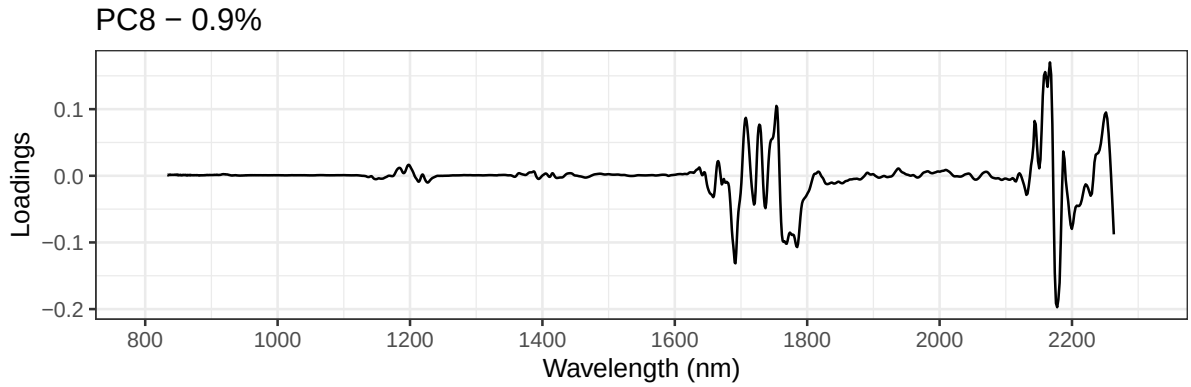


FIGURE E.15 – Loadings for PC8, which explains 0.9% of the variance.

List of publications and conferences

The research conducted as part of this thesis has contributed to scientific publications and conference presentations. Below, we provide a list of the published articles and conference contributions derived from this research.

- **Publications:**
  - A. Venegas-Reynoso, L. Giarracca-Mehl, M. Lacoue-Negre, B. Creton, C. Ruckebusch, and L. Duponchel. Identification of Key Molecular Features in Liquid Phase Autoxidation of Hydrocarbons. *Energy & Fuels*, 39(2):1192–1201, 2025. ISSN 0887-0624. doi: 10.1021/acs.energyfuels.4c04653.
  - A. Venegas-Reynoso, B. Creton, L. Giarracca-Mehl, M. Lacoue-Negre, C. Ruckebusch, and L. Duponchel. Oxidation Stability of Hydrocarbons: A Machine Learning-based Study. *Energy & Fuels* (Accepted).
- **Conferences:**
  - A. Venegas-Reynoso, L. Giarracca-Mehl, M. Lacoue-Negre, B. Creton, C. Ruckebusch, and L. Duponchel. Investigation of descriptors for the understanding and prediction of fluid oxidation stability with cheminformatics. *11th Colloquium Chemiometricum Mediterraneum* (CCM XI 2023), Padova, Italy. From 27 till 30 June 2023 (Poster).