THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Physics-Informed Data-Analysis Frameworks for Thermochemical Conversion Processes

Constraint-Aware Modeling for Steam Cracking of Polymeric Feedstocks in Dual Fluidized Beds

RENESTEBAN FORERO FRANCO

Department of Space, Earth and Environment

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2026

Physics-Informed Data-Analysis Frameworks for Thermochemical Conversion Processes

Constraint-Aware Modeling for Steam Cracking of Polymeric Feedstocks in Dual Fluidized Beds

ISBN 978-91-8103-385-4

DOI: https://doi.org/10.63959/chalmers.dt/5842

Acknowledgements, dedications, and similar personal statements in this thesis reflect the author's own views.

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5842
ISSN 0346-718X

Department of Space, Earth and Environment
Division of Energy Technology
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000

Printed by Chalmers Digitaltryck

Gothenburg, Sweden 2026

Physics-Informed Data-Analysis Frameworks for Thermochemical Conversion Processes
Constraint-Aware Modeling for Steam Cracking of Polymeric Feedstocks in Dual Fluidized Beds

RENESTEBAN FORERO FRANCO
Division of Energy Technology
Department of Space Earth and Environment
Chalmers University of Technology

Abstract
The escalating global production and consumption of plastics pose a significant environmental threat, demanding innovative waste management solutions. Thermochemical conversion via steam cracking offers a promising alternative to mechanical recycling, enabling the processing of highly heterogeneous plastic waste streams and recovery of carbon into monomeric species and valuable chemicals suitable for producing virgin-quality materials. The resulting product slate from steam cracking includes syngas, aliphatics, aromatics and soot, and is intrinsically linked to reactor conditions and feedstock's chemical characteristics. From a data analysis perspective, this distribution holds valuable information of the process that can be leveraged for instrument validation and estimation of unmeasured species, and relevant process variables. However, the high-dimensional, partially observed, and structurally diverse nature of the data demands robust, physically consistent analytical frameworks.

This thesis contributes to the theoretical and practical understanding of thermochemical conversion processes from a data analysis perspective. It introduces a constraint-aware methodology for developing data-driven models that can serve as tools for process optimization and to gain insights into the relationship between feedstock characteristics and process outcomes. The central idea is to merge first-principles constraints with compact statistical representations to build mathematical frameworks known as Constraint Networks (CN), enabling high-dimensional experimental systems to be transformed into low dimensional, tractable, and physically consistent models.

The research encompasses the development and validation of two complementary data-driven models. The Parametric System Model (PSM) transforms product species distributions into low-dimensional, physics-informed representations using discrete probability functions. By embedding conservation laws and topological constraints into a CN framework, the PSM enables physically meaningful estimation of unmeasured species, assessment of data quality, and validation of experimental setup. The Carbon Bond Group (CBG) model reduces both feedstock and product spaces into chemically meaningful vectors based on bond group environments. This dimensionality reduction allows steam cracking to be represented as a column-stochastic transformation between feed and products, facilitating cross-feed comparisons and enabling structure-based predictions through machine learning implementations. The models' development and validation were done using a pool of experimental data generated from steam cracking of various polymeric feedstocks under different operating conditions in a semi-industrial scale dual fluidized bed (DFB) reactor.

Overall, this work summarizes efforts to create generalizable data analysis frameworks aligned with physical principles for high-temperature thermochemical conversion systems. It contributes a scalable, interpretable, and constraint-centric modeling approach that supports the development of digital tools for process control and design, helping to pave the way for the future technology integration into circular economy strategies.

Keywords: Steam Cracking, Thermochemical Conversion, Plastic Waste Recycling, Fluidized Bed, Physics-Informed Modeling, Constraint Programming, Machine Learning, Data Reconciliation.

List of Publications Included in this Thesis

This thesis is based on the following papers, which can be found in the appendix of this work. The papers are referred to in the text using Roman numerals.

Paper I

Renesteban Forero-Franco, Teresa Berdugo-Vilches, Chahat Mandviwala, Martin Seemann, Henrik Thunman

Developing a Parametric System Model to describe the Product Distribution of Steam Pyrolysis in a Dual Fluidized Bed,

Fuel, Volume 348, 2023, 128518, ISSN 0016-2361, https://doi.org/10.1016/j.fuel.2023.128518

Paper II

Renesteban Forero-Franco, Isabel Cañete-Vela, Teresa Berdugo-Vilches, Judith González-Arias, Jelena Maric, Henrik Thunman, Martin Seemann,

Correlations between product distribution and feedstock composition in thermal cracking processes for mixed plastic waste,

Fuel, Volume 341, 2023, 127660, ISSN 0016-2361, https://doi.org/10.1016/j.fuel.2023.127660

Paper III

Chahat Mandviwala, Renesteban Forero Franco, Ivan Gogolev, Judith González-Arias, Teresa Berdugo Vilches, Isabel Cañete Cañete Vela, Henrik Thunman, Martin Seemann,

Method development and evaluation of product gas mixture from a semi-industrial scale fluidized bed steam cracker with GC-VUV,

Fuel Processing Technology, Volume 253, 2024, 108030, ISSN 0378-3820, https://doi.org/10.1016/j.fuproc.2023.108030

Paper IV:

Chahat Mandviwala, Renesteban Forero Franco, Teresa Berdugo Vilches, Ivan Gogolev, Judith González-Arias, Isabel Cañete Vela, Henrik Thunman, Martin Seemann

Steam cracking in a semi-industrial dual fluidized bed reactor: Tackling the challenges in thermochemical recycling of plastic waste

Chemical Engineering Journal, Volume 500, 2024, 156892, ISSN 1385-8947, https://doi.org/10.1016/j.cej.2024.156892

Paper V:

Renesteban Forero-Franco, Isabel Cañete-Vela, Teresa Berdugo-Vilches, Chahat Mandviwala, Nidia Díaz Perez, Ivan Gogolev, Henrik Thunman, Martin Seemann

Towards Sustainable textile waste management: Exploring valuable chemicals production through steam cracking in a dual fluidized bed

Fuel, Volume 397, 2025, 135731, ISSN 0016-2361, https://doi.org/10.1016/j.fuel.2025.135731.

Paper VI:

Renesteban Forero-Franco, Teresa Berdugo-Vilches, Chahat Mandviwala, Nidia Díaz Perez, Isabel Cañete-Vela, Henrik Thunman, Martin Seemann

Physics-Informed Framework for Predictive Modeling of Product Distributions in Steam Cracking of Heterogeneous Polymeric Waste

Submitted for publication.

Contributions Report:

Paper I. Renesteban Forero Franco is the corresponding author, with responsibility for the model conceptualization, data curation, formal analysis and writing of the paper; Teresa Verdugo contributed with editing, ideas and discussion regarding the manuscript, and planning of the experimental work. Chahat Mandviwala contributed to the execution of the experimental work; Martin Seemann contributed with ideas and discussion during the preparation of the manuscript; Henrik Thunman contributed with ideas and discussion regarding the model formulation and preparation of the manuscript.

Paper II. Renesteban Forero Franco is the corresponding author, with responsibility for the model conceptualization, data curation, formal analysis and writing of the paper; Isabel Cañete Vela contributed with the planning and execution of the experimental work, as well as preliminary ideas and editing for the final version of the manuscript; Teresa Verdugo contributed with editing, ideas and discussion regarding the manuscript and the planning of the experimental work; Judith Gonzales Arias contributed with ideas and evaluation of the experimental data; Jelena Maric contributed to the execution of the experimental work; Henrik Thunman contributed with discussions and editing of the manuscript. Martin Seemann contributed ideas and discussions regarding the model formulation and preparation of the manuscript.

Paper III. Renesteban Forero Franco is a corresponding co-author, with responsibility for data curation, conceptualization and editing, as well as the planning and execution of the experimental work; Chahat Mandviwala is a corresponding co-author with responsibility for the writing and conceptualization of the manuscript, as well as the planning and execution of the experimental work;; Ivan Gogolev, Judith Gonzales Arias and Isabel Cañete Vela contributed to the execution of the experimental work; Teresa Verdugo contributed with editing of the manuscript and planning of the experimental work; Henrik Thunman and Martin Seemann contributed with supervision, ideas, discussions and editing of the manuscript.

Paper IV. Renesteban Forero Franco is a corresponding co-author, with responsibility for data curation, conceptualization and editing, as well as the planning and execution of the experimental work; Chahat Mandviwala is a corresponding co-author with responsibility for the writing and conceptualization of the manuscript, data curation as well as the planning and execution of the experimental work; Teresa Verdugo contributed with editing, ideas and discussion regarding the manuscript, and planning of the experimental work; Ivan Gogolev, Judith Gonzales Arias and Isabel Cañete Vela contributed to the execution of the experimental work; Henrik Thunman and Martin Seemann contributed with supervision, ideas, discussions and editing of the manuscript.

Paper V. Renesteban Forero Franco is the corresponding author, with responsibility for the model conceptualization, data curation, formal analysis and writing of the manuscript; Isabel Cañete Vela contributed with the planning and execution of the experimental work, data curation, conceptualization and editing for the final version of the manuscript; Teresa Verdugo contributed with editing, ideas and discussion regarding the manuscript and the planning of the experimental work; Chahat Mandviwala, Ivan Gogolev, and Nidia Diaz Perez contributed to the execution of the experimental work, ideas and evaluation of the experimental data; Henrik Thunman contributed with discussions and editing of the manuscript. Martin Seemann contributed with editing, ideas and discussions regarding the conceptualization and preparation of the manuscript.

Paper VI. Renesteban Forero Franco is the corresponding author, with responsibility for the model conceptualization, data curation, formal analysis and writing of the paper; Teresa Verdugo contributed with editing, ideas and discussion regarding the manuscript and the planning of the experimental work; Chahat Mandviwala contributed with editing, ideas and discussion of manuscript, planning, execution and evaluation of the experimental work and data, Nidia Diaz Perez and Isabel Cañete Vela contributed to the execution of the experimental work, ideas and evaluation of the experimental data; Henrik Thunman contributed with discussions, ideas and editing of the manuscript. Martin Seemann contributed with editing, ideas and discussions regarding the conceptualization and preparation of the manuscript.

This thesis is also based on parts of the licentiate thesis published by the author as part of the evaluation of his doctoral studies:

Renesteban Forero Franco, Data-Driven Models and Correlation Strategies for Thermal Cracking of Polymeric Feedstocks in Dual Fluidized Beds, 2024

Other Related Publications not included in this thesis:

Judith González-Arias, Renesteban Forero-Franco, Chahat Mandviwala, Martin Seemann,

Steam gasification as a viable solution for converting single-use medical items into chemical building blocks with high yields for the plastic industry,

Resources, Conservation and Recycling, Volume 201, 2024, 107342, ISSN 0921-3449, https://doi.org/10.1016/j.resconrec.2023.107342.

Ivan Gogolev, Nidia Diaz Perez, Chahat Mandviwala, Renesteban Forero Franco, Ann-Christine Johansson, André Selander, Martin Seemann,

Thermochemical recycling of mixed plastic wastes through pyrolysis and steam cracking – Assessment of centralized vs. Decentralized approaches,

Thermal Science and Engineering Progress, Volume 62, 2025, 103558, ISSN 2451-9049, https://doi.org/10.1016/j.tsep.2025.103558.

Acknowledgments

Reaching the end of this five-year journey has only been possible thanks to the many people who supported, guided, and believed in me throughout the process.

First and foremost, I would like to express my sincere gratitude to my supervisors and seniors: Martin Seemann, Henrik Thunman, and Teresa Berdugo.

Martin, your supervision and our discussions were invaluable. They continuously helped me reorient my research path whenever I felt uncertain and shaped the direction of this PhD in profound ways. Whenever I felt undecided, I always knew I could count on you to talk things through and gain the perspective I needed to set a clear direction. I have always enjoyed your sense of humor, which has brightened my days many times in our casual conversations.

Henrik, your ability to challenge me with new and inspiring ideas pushed me to think beyond the obvious and to explore amazing perspectives that turned my thinking upside down in the best way. Thank you for always being available whenever I needed guidance and new angles during my research.

Teresa, your guidance, kindness, and unwavering availability were a constant anchor. Your thoughtful feedback and careful reviews elevated the quality of my work, and your support helped me maintain focus even in the most demanding phases.

I would also like to thank my teammates Chahat, Nidia, Isabel, Tharun, Eliette, Judit, Adam and Iacopo for creating a friendly and motivating work environment in the group, as well as Jessica, Johannes and Rustan for their invaluable help on our experimental campaigns. Thanks a lot Johannes for all your patience and dedication to get the last two degrees I always needed in the reactor's temperature.

A special thanks to Chahat, my office mate throughout the entire PhD. As I once told you, your friendship has great value to me on more levels than one. I deeply treasure all our interactions, conversations, and shared experiences; they were a significant part of my daily life. Your knowledge in the field gave me confidence many times to take bold approaches, and you were always open to sharing that expertise and offering help without restraint. I am also grateful to Nidia for being such a wonderful teammate. Whenever we met in your office or mine, our conversations brought laughs and fresh energy to my days at several moments.

I want to thank my father, who taught me the principles of discipline, dedication, and moral integrity that continue to guide my life. His teachings have been a source of strength and direction throughout this journey.

My deepest and most indescribable gratitude goes to my wonderful wife Ivonne. You have been my partner, my friend, my confidant in every adventure of this journey, through its highs and lows. Your love, care, patience, and unwavering encouragement have been my greatest support. Your presence in my life each day has meant the world to me, and this achievement is as much yours as it is mine.

Finally, I owe all of this to my God and Lord Jesus Christ who gave me the faith, strength, inspiration and most firm foundation to keep myself on my feet every day and take this journey to the end.

“dubito ergo cogito, cogito ergo sum”

Table of Contents

This page intentionally left blank

1. Introduction

1.1. Background

Plastics have revolutionized modern life due to their light weight, versatility, and low cost, but their rapid proliferation has created a major global environmental challenge. Global production of plastics exceeded 400 million metric tons (Mt) in Year 2022 and projected to double by year 2050 if the current trend continues, while the effective recycling rate reaches around 9% [1,2] . However, the plastics economy is far from circular. In the same year, around 268 Mt of plastic waste was produced worldwide, but only 38Mt was effectively recycled, which accounts for less than 10% of the primary worldwide production. The remaining waste ends up landfilled (∼50%), incinerated or mismanaged [2,3].

In response to these problems, current recycling systems include mechanical recycling as the most common approach. This process involves the re-melting and shaping of plastics for reuse. However, it faces several limitations, due to strict requirements in terms of homogeneity and low feedstock contamination. Multiple sorting steps are typically required when dealing with heterogeneous plastic feedstocks, which reduces the overall energy efficiency of the process. Additionally, the quality of the material progressively degrades with each recycling cycle [4,5]. As a result, a significant portion of the collected plastic waste ends up as rejected fraction, where carbon atoms may only be recycled through more-general routes, such as combustion combined with carbon capture and utilization (CCU) technologies [6,7], but come with an unavoidable increase in energy demand.

Moreover, mechanical recycling inherently preserves the material characteristics tailored for current market demands. This means that the additives, polymer structures, and functionalities embedded in today's plastics are perpetuated through each recycling cycle. This limits the ability to adapt to future needs, where different performance characteristics or environmental considerations may become more relevant for the materials. Thus, while mechanically recycling methods may be more energy-efficient, they risk locking material's design into the specifications of the past.

Alternatively, chemical recycling offers a more flexible and general recycling approach, capable of processing heterogeneous plastic waste streams to yield valuable chemicals. This method focuses on transforming back the feedstock's molecules into their fundamental chemical building blocks, monomers, which can be used as raw materials for manufacturing virgin-quality plastics and other high-quality products. By doing so, this recycling route provides the adaptability needed to meet future material demands. Among the most prominent techniques are solvolysis and pyrolysis. Solvolysis involves the chemical degradation of polymers using solvents, while pyrolysis is a thermochemical process which relies on high temperatures in an oxygen-free environment to break down the polymer chains through free radical reactions. In today's global context, chemical recycling can be seen as complementary to mechanical recycling, with the potential to process mixed or contaminated plastics at industrial scale [8,9].

Among the thermochemical conversion technologies, steam cracking is a process of special interest due to its current application for monomer production from Naphtha. This process involves subjecting carbon-based feedstocks to high temperatures, typically around 800°C, in the presence of steam, which acts as a quasi-inert dilution agent. This pyrolysis-based conversion process leads to the generation of a gas mixture, technically known as *producer gas*, which contains a variety of chemical species, including H_2, CO, CO_2, and various hydrocarbons in the forms of paraffins, olefinic monomers, other olefins and aromatics. This applies to any kind of carbon-based feedstock but the species distribution of the producer gas is directly linked to the cracking reactor conditions and the chemical characteristics of the feedstock. Since it can process different kinds of heterogenous waste feedstocks, the technology falls within the general recycling category but with the advantage of minimizing the thermodynamic penalty associated with the combustion-based recycling route [8]. This conversion process constitutes the core of the research application presented in this work.

The endothermic nature of pyrolysis-based reactions requires continuous heat input to the reaction environment, where a steep heating gradient and a controlled maximum temperature benefits the product distribution towards monomer recovery. Therefore, the reactor type and design play pivotal roles in polymer conversion and the resulting product species distribution. Dual Fluidized Beds (DFBs) fulfill the abovementioned requirements due to their allothermal heat supply, efficient heat transfer capabilities, and effective mixing characteristics. In a DFB system, the reactor consists of a combustor and a pyrolysis unit, with a heated sand bed operating in a circulating or bubbling fluidized regime, serving as a heat carrier between both units (See Figure 1).

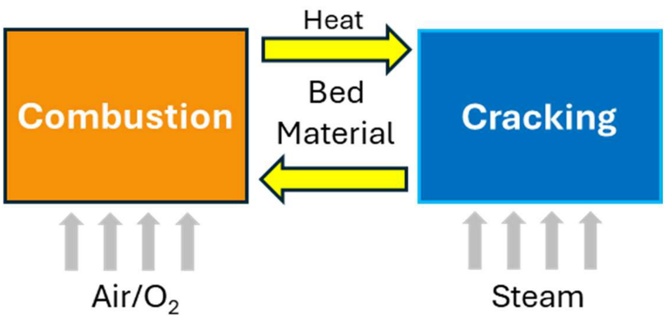


Figure 1. DFB reactor system schematic

This configuration offers flexibility in fine-tuning reaction conditions to favor desired product distributions. DFBs also offer high versatility in terms of feedstock size and state, enhancing the robustness of the conversion process when handling diverse input materials. Integration of such systems within petrochemical facilities has been explored with the goal of recycling 100% of the carbon present in plastic waste streams [8]. As such, this technology presents a promising alternative to fossil-based monomer production for the chemical industry, and it is the basis of the experimental setup used in this work.

From a modeling standpoint, capturing the behavior of this system presents unique challenges. Kinetic models and purely data-driven methods are commonly used approaches, but both present clear limitations when applied to this context. Detailed deterministic kinetics is the traditional route to model the behavior of the cracking process at the free-radical reactions level. However, these models grow rapidly in dimensionality and complexity as more reactions, intermediate species, and parameters must be considered, especially when dealing with heterogeneous feeds and the heat transfer dynamics within the DFB operation. This complexity makes model structuring, parameter identifiability, and practical implementation demanding in both time and computational resources. On the other hand, advanced purely data-driven models, such as Neural Networks (NN), can capture trends and deliver fast predictions, yet they act as black-box models often disregarding conservation laws and physical plausibility, which greatly limits interpretability and extrapolation [10]. To address these limitations, recent efforts have focused on intermediate approaches that combine collected data with domain knowledge to develop compact, physics-aware data analysis models. These models incorporate information such as conservation laws, chemical feasibility, and system regime characteristics to provide structural guidance for data interpretation, inference, and prediction [11–13]. This is the general stance explored in the modelling frameworks developed in the present work.

Beyond the technological capabilities of steam cracking in DFB systems, a central challenge remains, which is common in many complex process systems: the data obtained from their operation are high-dimensional, often partially observed, and shaped by heterogeneity in both input feed and operating conditions. In the steam cracking context, mixed polymer streams, variable ash and heteroatom contents, and shifting reactor hydrodynamics all influence the measured product species slate. It is also common to find data with gaps, such as incomplete species coverage, biases from calibration limits, or drift between parallel analytical setups. This complexity calls for analysis methods that go beyond purely data-driven black-box models. Instead, they must be system-aware containing domain knowledge, so that predictions, inferences, and data quality checks retain physical meaning and remain faithful to conservation laws and the underlying chemistry of the process. These requirements align with and motivate the use of reduced complexity, structure informed modelling strategies that have been explored in related petrochemical and thermochemical applications, forming the conceptual starting point for the approaches developed in this thesis.

Modelling strategies that reduce chemical complexity into compact, domain-aware representations have long been used in the petrochemical and thermochemical sectors, particularly as a way to manage high-dimensional product slates without resorting to fully deterministic kinetics. In Fischer–Tropsch synthesis and polyethylene pyrolysis, for example, the Flory-Schulz distribution and similars (e.g. Anderson–Schulz–Flory distribution) provides a parametric carbon-number description of paraffinic products, enabling characterization of broad hydrocarbon slates under varying catalysts and operating conditions [14,15]. Similarly, in conventional steam cracking and in molecular-reconstruction methods for petroleum fractions, carbon-number distributions and pseudo-component lumping are used to relate feed properties to product trends and to support extrapolation across operating regimes [16–19]. In parallel, functional-group–based formulations developed for real-fuel combustion and lumped kinetic models used in biomass and coal pyrolysis (e.g. FGMech models) demonstrate how reduced structural descriptors can capture aggregate conversion behavior across diverse feedstocks and simplify until certain degree full deterministic kinetic calculations [20,21]. On the data-quality side, data-reconciliation frameworks that enforce elemental balances and correct measurement inconsistencies are well-established tools in process systems engineering, particularly for thermochemical systems sensitive to analytical uncertainty [22–24]. However, these approaches generally assume well-characterized and compositionally stable feeds under controlled operating conditions, assumptions that do not hold for heterogeneous polymeric waste streams, where feed variability, incomplete characterization, and measurement gaps become intrinsic modelling challenges.

Building on the unifying idea behind these conventional models, particularly the use of structural descriptors, dimensionality reduction, and formulations that enforce mass and elemental balances to interpret complex hydrocarbon systems, this work adapts such principles to the specific challenges posed by heterogeneous polymeric feedstocks under steam cracking. Rather than reproducing the detailed kinetic fidelity of functional group mechanisms or microkinetic models, the objective here is to construct domain aware, reduced order representations that remain compatible with the uncertainty and partial observability present in the thermochemical conversion of polymer derived feeds. To that end, the methodology formalizes conservation laws, chemically meaningful reductions of state, and related invariants into a coherent, constraint centric analytical framework. Constraint Networks (CNs) provide the mathematical umbrella for encoding these rules as feasible sets, enabling the formulation and structuring of problems whose solutions support descriptive, predictive, and prescriptive analytics without violating first principles.

Two complementary models are developed in this work to operationalize this framework. First, the Parametric System Model (PSM) addresses the complexity and partial observability of product distributions by embedding a structured set of constraints within a reduced parametric representation of the cracking outputs. It transforms high dimensional instances from structure indexed product group sets (e.g, paraffins and olefins resolved by carbon number) into low dimensional distribution functions that satisfy conservation laws as well as statistical and topological conditions. Solving this constrained system yields a closed form, low parameter representation of the product slate that remains physically plausible and can be used both as a predictive tool and as a mechanism to assess data quality and internal consistency.

Second, the Carbon Bond Group (CBG) model extends the complexity reduction to both the input and output spaces by identifying common chemical environments that can be used to characterize and correlate heterogeneous feedstocks and the resulting conversion products. By expressing feeds and products through comparable structural descriptors, the model compresses the high dimensional variability of polymer mixtures and product species into interpretable low dimensional vectors. Steam cracking is then represented as a mathematical transformation between comparable spaces that maps feed bond environments to product families through a compact operator that encodes the conversion process. This relational structure supports cross-feed comparisons, enables the identification of systematic trends, and provides additional physics informed constraints that can be integrated with other data analysis methods to create tools useful for real-time operation and control.

Altogether, the motivation of this thesis is to shift from ad-hoc, case-dependent data interpretation toward a generalizable and constraint aware analytical perspective for complex thermochemical systems. While the specific demonstrations focus on DFB steam cracking of polymer rich waste, the underlying CN based approach is generic, as any conversion system that admits a consistent carbon class or analogous basis for defining variables and hard constraints can adopt the same scaffolding. By embedding conservation laws, chemically meaningful dimensionality reduction, and regime-dependent mappings into the analytical methodology, the approach aims to construct models with the structural consistency required to: (i) improve confidence in experimental data interpretation through internal consistency checks, (ii) enable predictive use of heterogeneous datasets for quick scenario exploration, and (iii) support the integration of advanced methods that enable online characterization of the process outputs as actionable feedback in adaptative process control systems. These elements together provide a principled foundation for creating compact, interpretable, and operationally useful models that remain adaptable to fit the complexity of thermochemical conversion systems.

In doing so, this work contributes to broader efforts to transition plastics toward a circular value chain. By applying physics informed data analysis and machine learning methods to DFB steam cracking, the thesis delivers reproducible and auditable modelling frameworks that link feedstock characteristics to product slates, enhance data quality and inference, and provide actionable insights to support the integration of this technology into circular plastics lifecycles.

1.2. Aim and Scope

The overall aim of this work is to derive and validate a generalized data analysis methodology based on physical and statistical constraints to evaluate highly complex systems with heterogenous and complex interrelated inputs and outputs variables. The models derived are intended to enable robust data quality assessment, predictive modeling, and estimation of key unknown quantities. The methodology is applied to assess and generalize experimental results from steam cracking of heterogeneous carbonaceous feedstocks in dual fluidized bed reactors for recovery of monomers and other valuable chemicals.

The present work compiles research results of investigations with different polymer-based materials and conditions performed on a semi-industrial DFB steam cracker plant. From the collected pool of data, it is explored the mathematical framing and implementation of physics-informed data analysis models to improve the quality of characterization outcomes as well as evaluate the possibilities of create high performing models of the reactor as tools towards processing control and operation.

To accomplish the aim, the following research goals are pursued:

- Identification of the generic mathematical framework to build models that allow to incorporate domain-knowledge into generic data analysis
- Develop a physically and statistically-constrained data analysis framework that can be used to assess the characterization outcomes downstream the thermochemical conversion process, to identify measurement inconsistencies, assess data quality, and give meaningful estimations of unmeasured product species.
- Develop experimental setup methodology to obtain a comprehensive characterization of the carbon from the process to test and validate the models.
- Explore how the chemical structural characteristics of a feedstock can be related to the product species distribution in order to develop a common framework to interrelate inputs and outputs of complex heterogeneous feedstocks at a tractable level
- Integration of Machine Learning techniques to develop physics-informed models that interrelate inputs and outputs and serve as digital twin like models of the steam cracking reactor. Such integration can be used towards data analysis to enhance knowledge of process outcomes, predictive capabilities, and operational flexibility in dual fluidized bed steam cracking of polymeric waste feedstocks

1.3. Contributions to this Thesis

This thesis comprises the main developments and findings from six scientific papers. The research's exploration has been conducted at both theoretical and experimental levels. This document presents the equations and mathematical foundations of the developed models, as well as their validation using real industrial-scale data. The summary description of the papers contributions is described below:

Paper I introduces the concept, mathematical background and experimental validation of the custom-made model, called the Parametric System Model (PSM), used for data quality assessment of product species distributions is presented.

Paper II focuses on understanding how the product distribution is influenced by the feedstock polymeric composition when cracking highly heterogeneous mixtures in a DFB. It presents a novel carbon bond group (CBG) classification framework that was developed to explore correlations between the products and the feedstock chemical structures.

Paper III showcases experimental methods and characterization setups developed to get a quick and comprehensive carbon balance from the cracking gas product, with particular focus on GC with Vacuum Ultraviolet Spectroscopy (GC-VUV) as the characterization method.

Paper IV evaluates DFB steam cracking on an operational window as a flexible technology for producing valuable chemicals from both conventional and waste-derived feeds. It benchmarks product patterns and process fingerprints across representative feedstock types (Naphtha and polymeric wastes), confirming the consistent formation of light olefins and monoaromatics with steam-cracking chemistry, while identifying DFB-specific signatures. The work also outlines challenges relevant to thermochemical recycling and highlights practical operating considerations and data-assessment strategies that make the reactor and analytics suitable for model validation and scale-relevant evaluation.

Paper V explores the challenge of mixed textile wastes by investigating DFB steam cracking of polyester- and cellulose-rich waste streams to examine composition–outcome relationships and characteristic conversion indices behavior. It evaluates the distinct shifts in product distributions driven by feed chemistry and temperature regime, showing how oxygen-rich structures influence target products and how their processing via steam cracking can contribute to circular strategies.

Paper VI extends the CBG framework into a predictive, physics-informed model for steam cracking of heterogeneous polymeric waste. The reactor is formulated as a trainable mapping between feed and product compositional spaces defined in the CBG basis, embedding structural constraints and domain knowledge by design. Temperature dependence is incorporated into the conversion operator to evaluate operating conditions effects. The model demonstrates strong predictive performance and robust generalization across diverse feedstocks, while revealing chemically interpretable trends in carbon redistribution. The generalized CBG formulation thus provides a reduced-order, scalable surrogate with potential for digital twin development of thermochemical recycling systems.

1.4. Thesis Outline

This work is organized as follows. In chapter 1 the problem context in introduced, as well as motivation, aims, scope, and the included papers' contributions. Chapter 2 establishes the data-analytics foundation by formalizing the Constraint Networks (CN) as the mathematical framework for embedding physics and other domain-knowledge constraints into data-driven models. Chapter 3 reviews the thermochemical background of steam cracking, highlighting mechanisms and physical limits that motivate the conservation, topology, and inequality constraints used later. Chapter 4 develops the modelling toolkit: the Parametric System Model (PSM) to constrain and evaluate product species distribution; description of the Carbon-Bond-Group (CBG) to link feedstock structural characteristics to products and conceptualization into low-dimensional spaces mapping, and machine learning integration to create composition-estimation formulation with physics-aware regularization. Chapter 5 describes the reactor's experimental setup, sampling and characterization strategy and evaluated conditions and feedstocks. Chapter 6 presents results and discussion, covering PSM fits, diagnostics and validation checks, as well as the CBG model proof-of-concept results across different heterogeneous feeds. Chapter 7 presents an outlook of CBG model extensions towards reactor-dependent generalization, and dimensionality increase to capture residual variance, and the PSM and CBG models integration pathways toward adaptative control and process optimization. Chapter 8 presents summary of findings, consolidated contributions, and implications for integrating DFB steam cracking into circular processing strategies. At the end of the document, the reader can find the list of abbreviatures and an appendix dictionary with definitions of special and technical terminology and concepts used in this work.

2. Data analysis concepts and physics informed models

2.1. Data Analysis fundamentals and Constraint Networks (CN) problems overview

The techniques employed in data analysis span descriptive, predictive, and prescriptive analytics, each contributing to the understanding of the data generated out of a system and the application of the knowledge obtained out of it.

In descriptive analytics, the analysis provides a preliminary assessment of the data by summarizing their key characteristics using statistical metrics (such as mean, median, mode, variance, and standard deviation) and visualization techniques, to detect trends and statistical distributions. Predictive analytics develop models to estimate outcomes, anticipate changes and future trends based on historical data. Finally, prescriptive analytics integrates descriptive and predictive outcomes to provide insights into patterns and future outcomes to make data-driven decisions and define courses of action based on optimization models. To effectively implement prescriptive or even predictive analytics, it is essential to clearly define a structure of conditions or limitations that the data-derived solutions must satisfy. This structured approach aligns closely with Constraint Networks, a mathematical concept that provides an underlying representation to define problems governed by specific rules and restrictions that require a solution.

Constraint Networks (CN) are a fundamental framework in artificial intelligence to declare problems defined by a set of variables that must assign values subject to constraints [25]. As a mathematical framework, a CN instance is formulated over three fundamental sets of elements: variables (V), domains (D), and constraints (C), forming a triple $\langle V, D, C \rangle$. Each element is formally defined as:

- Variables: $V = \{V_1, \dots, V_n\}$. Represent the set of unknowns or elements to be solved.
- Domains: $D = \{D_1, \dots, D_n\}$. Each variable V_i has an associated non-empty domain D_i, which includes all possible values the variable can assume.
- Constraints: $C = \{C_1, \dots, C_m\}$. Rules or relationships between variables that restrict the combinations of values the variables can take simultaneously. Constraints can be unary, binary, or higher-order k-ary, depending on whether they involve one, two, or k multiple variables. Formally, each constraint C_j is a pair $C_j = (scope_j, relation_j)$. The $scope_j$ is a subset of k variables from V, i.e., $scope_j = \{V_{i_1}, \dots, V_{i_k}\} \subseteq V$. The $relation_j$, is a defined subset of the cartesian product of the domains of the variables in $scope_j$, i.e., $relation_j \subseteq D_{i_1} \times \dots \times D_{i_k}$. This subset specifies the allowed combinations of values for those variables

Example:
Consider methane steam reforming under simplified stoichiometric conditions as quick example of the concept's application: $CH_4 + H_2O \leftrightarrow CO + 3H_2$. Assume 1 mol of CH_4 and 1 mol of H_2O are fed to the reactor, and only the primary reforming reaction is considered. In a CN formulation, the variables are the amounts of key products, such as n_{CO}, $n_{\mathrm{H_2}}$, and optionally $n_{\mathrm{CO_2}}$ if the water-gas shift is considered. Each variable has a domain, for example $n_{\mathrm{CO}} \in [0,1]$ mol and $n_{\mathrm{H_2}} \in [0,3]$ mol, reflecting stoichiometric limits. The constraints enforce physical laws: carbon balance ($n_{\mathrm{CO}} + n_{\mathrm{CO_2}} = 1$), hydrogen balance, and non-negativity. Additional constraints can represent performance targets, such as $n_{\mathrm{H_2}} \geq 2.5$ mol for high hydrogen yield. Solving the CN means finding a set of product quantities that satisfies all these rules. For instance, ($n_{\mathrm{CO}} = 1, n_{\mathrm{H_2}} = 3, n_{\mathrm{CO_2}} = 0$) is a valid solution under ideal conversion conditions. This simple example illustrates how CNs translate chemical feasibility and conservation principles into a structured problem-solving framework.

In general, problems described as CNs can be categorized based on certain characteristics which determine the solution method to follow [25–29]:

I. *Based on Constraint Arity:*

a) Unary or Binary: Simplest form of CN, where all constraints involve exactly one (unary) or two (binary) variables.

b) k-ary (high order): Also called non-binary CN. Involve constraints that apply to three or more variables simultaneously. In general, many problems can be converted into an equivalent binary CN, making this a common focus for theoretical analysis. and is one of the strategies explored in this work.

II. *Based on Domain Type:*

a) Discrete Domain: A common type of CN, where each variable has a countably finite or infinite domain (e.g. a subset of natural numbers).

b) Continuous Domain: Here, variables can take on an uncountably infinite number of values (e.g., real numbers within a range). These often require different solution techniques, sometimes involving specialized numerical methods rather than discrete search.

III. *Based on Problem Dynamics:*

a) Static CN: The variables, domains, and constraints are fixed and do not change during the problem-solving process. This is the standard formulation.

b) Dynamic CN: The problem formulation can change over time, meaning variables or constraints might be added, removed, or modified during the search for a solution. Real-time scheduling or resource reallocation in changing environments are examples. These require algorithms that can adapt and often reuse information from previous problem states.

IV. *Based on Solution Requirements:*

a) Satisfactory CNs: The goal is simply to find any assignment that satisfies all constraints, or to determine that no such assignment exists. An assignment found will not necessarily correspond to the optimal system's solution.

b) Optimization CNs: In these problems, it might be impossible to satisfy all constraints simultaneously. The goal is then to find an assignment that optimizes a specific requirement, encoded in a function, which satisfies the maximum possible number of constraints.

Based on these characteristics, two fundamental problem classes are the most commonly distinguished in CNs: Constraint Satisfaction Problems (CSP) and Constraint Optimization Problems (COP). A classic CSP operates on variables with finite, discrete domains and its solution requirement is strictly of the satisfactory type. Formally, the goal in a CSP is to find a complete assignment A of values $v_i \in D_i$, one value v_i to each and all variables V_i, such that all the constraints are satisfied. Such an assignment corresponds to a CSP's solution. If no such assignment exists, the CSP has no solution. On the other hand, a classic COP operates in the discrete and continuous domain and takes an optimization-based solution requirement, i.e., it seeks to find the optimal solution based on an objective function and the constraints imposed. Thus, a COP can be seen as a generalization of a CSP [25–27].

In this work's study case, a complex system such as the thermochemical conversion of plastic waste through steam cracking will be viewed under the framework of static k-ary CNs with continuous domain and over-constrained requirements. The models' constructs will be developed further in the next sections.

In general, solving CN instances typically involves various methods, often combined for efficiency, and fall within the scope of the so-called mathematical optimization and constraint programming fields [27,30]. Systematic Search Methods, Local Search and Metaheuristics, Convex Optimization and First-Order Solvers are common examples of solution approaches [25,31]. Systematic search methods, typically used in CSPs, are used for relative small CN systems and the most common method is known as backtracking search, which incrementally assigns values to variables and backtracks when an inconsistency (a violation of a constraint) is detected. To improve efficiency, common techniques used are: *Heuristics* for variable and value ordering to assign next, and *Constraint Propagation* that prunes the domains of unassigned variables by eliminating values that cannot possibly be part of any solution.

Larger or more complex CNs, where systematic search might be too computationally expensive, or unfeasible, such as in COPs, local search algorithms and metaheuristics are often employed [26,32]. These methods typically work by starting with a complete (though possibly inconsistent) assignment and iteratively trying to improve it by making small, "local" changes.

Local Search Algorithms explore the solution space by moving from one candidate solution to a "neighboring" solution that is "better" according to some evaluation function (e.g., minimizing the number of violated constraints). Common examples include Hill Climbing, which moves to the neighbor assignment that improves the objective [25]. A key limitation of simple local search is its tendency to get trapped in local optima.

On the other hand, Metaheuristics are higher-level algorithmic frameworks that guide underlying local search procedures to overcome their limitations, particularly the problem of being trapped by local optima. They provide a balance between exploration (diversification), which involves looking for new, unvisited regions of the search space, and exploitation (intensification), which focuses on refining promising areas.

Metaheuristics often incorporate stochastic (random) elements to facilitate exploration specially when dealing with high dimensions and multi-modal (multiple local optima) solution spaces. Examples include Simulated Annealing which is inspired by the annealing process in metallurgy, where occasional "worse" moves are accepted with a probability that decreases over time (temperature), in order to escape local optima regions [33]. Another metaheuristic example is Particle Swarm Optimization (PSO), a highly robust population-based method inspired by the social behavior of bird flocking. The algorithm maintains a "swarm" of candidate solutions (particles) that move through the search space, influenced by their own and the entire swarm's best-found positions [34].

Overall, a CN instance provides a concise declarative framework for modelling problems where decisions are inter-dependent and must satisfy explicit rules. By formalizing a problem as a set of variables, their associated domains, and the constraints that govern admissible combinations, CNs enable systematic exploration of the solution space and supports logical space deduction as the number of constraint increases. Intuitively, each additional constraint can be seen as a boundary that excludes infeasible assignments, progressively shrinking the hyperspace of possibilities (see Figure 2). Therefore, the tighter the feasible region is, the greater the confidence that any assignment that survives the filtering corresponds to a consistent rule-based solution to the system.

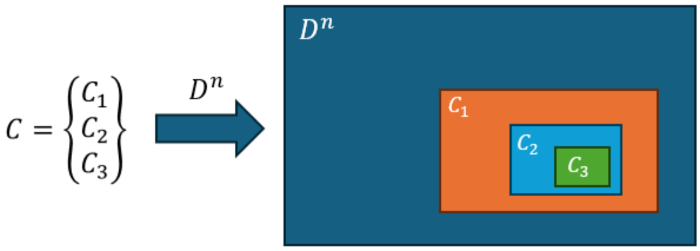


Figure 2. Representation of the effect of constraints on the solution hyperspace.

2.2. Constraint Types and Invariants

As discussed before, constraints define the relationships among variables that must be satisfied for a CN's solution to be valid. Constraints effectively shape the structure of the solution space, eliminating combinations that violate problem-specific logic or requirements. Depending on the nature of their enforcement, constraints can be grouped into two categories: Hard (C_H) and soft (C_S) constraints [27,30].

i. *Hard Constraints*

These constraints must always be satisfied. Any assignment that violates a hard constraint is considered invalid. Formally, a hard constraint $C_j \in C_H$ is satisfied only if the projection of the assignment $A = \{V_1 \leftarrow v_1, \dots, V_n \leftarrow v_n\}$[1] onto $scope_j$ satisfy $relation_j$. Therefore, any valid assignment (solution) of a CN instance implies that:

$$proj_{scope_j}(A) = \{v_{i_1}, \dots, v_{i_k}\} \in relation_j, \forall C_j \in C_H \tag{1}$$

In other words, the set of assigned values for the variables in $scope_j$ must be an element of $relation_j$. For context, recall that $scope_j$ corresponds to the set of k variables $\{V_{i_1}, \dots, V_{i_k}\}$ participating in constraint j. Therefore $proj_{scop\ j}(A)$ corresponds to the projection of that set into the assignment A (which is the set of the assigned values to all variables in the CSP). In other words, it is a set that only contains the values from variables that were part in the constraint j scope, i.e., $\{v_{i_1}, \dots, v_{i_k}\}$. Hence, Equation (1) tells that such filtered set of values must be within the domain

[1] The notation $V_1 \leftarrow v_1$ indicates the value v_1 is assigned to variable V_1, in other words, $A = \{v_1, \dots, v_n\}$

imposed by the hard constraint defined as $relation_j$, to be considered as valid. These constraints represent the core of a CSP instance, which can be defined as a CN of the form $\langle V, D, C_H \rangle$.

For computational applications, logical-based constraints can be derived as functions of the assignment $\{v_{i_1}, \dots, v_{i_k}\} = v_j$. Thus, in practice a hard constraint is satisfied only if $C_{logic}(v_j) = true$.

ii. Soft Constraints

These constraints express preferences rather than strict rules. Violating them doesn't invalidate a solution but may incur a cost. A soft constraint $Cj \in C_S$ is defined as a pair $C_j = (scope_j, \phi_j)$, where $\phi_j(v_j)$ is a scalar function $\phi_j: D_{i_1} \times \dots \times D_{i_k} \to \mathbb{R} \geq 0$ corresponding to the penalty or cost of violation of the constraint C_j, given the values v_j assigned to the variables in $scope_j$.

Therefore, soft constraints define a scoring function over assignments and are often integrated into an objective function $f: D^n \to \mathbb{R}$, where the goal is to find the assignment A that minimizes the function (also known as loss function):

$$\min_A f(A) + \lambda \cdot \sum_{j=1}^{m} \phi_j(v_j) \tag{2}$$

with λ controlling the trade-off between optimization goals and constraint violations ($\sum \phi_j$). This formulation constitutes the backbone of a COP instance which can be seen as an augmented CN to a 4-tuple in the form $\langle V, D, C_H, C_S \rangle$ [26]. Therefore, a classic COP's goal is to find a solution that satisfies all hard constraints while minimizing the total cost from violating soft constraints.

Invariants as Structural Constraints

In systems that undergo transformations over time or under certain operations, an important mathematical property to look after among the objects participating in the process are the invariants. An invariant is a quantity or relation that remains unchanged under transformations or evolution of the system's state. Bringing this concept to a CN, invariants can translate into hard constraints that reflect intrinsic properties of the system being modeled and can play a particular foundational role, as will be seen in the next sections of this work.

In a CN framework, let the system's state be described by the assignment $A = \{v_1, .., v_n\} = v$, with $V \leftarrow v$, and let $M: D^n \to D^n$ represent a transformation or process (e.g., a timestep in a simulation or a chemical conversion step) from $v \to v'$, with $v' = M(v)$ and $V' \leftarrow v'$. A function $I: D^n \to \mathbb{R}$ is an invariant if:

$$I(v) = I(M(v)), \forall v \in D^n \tag{3}$$

That is, I remains constant even as the system evolves ($I(v) = constant$). Thus, a constraint from an invariant will be defined as:

$$\begin{gathered} scope_j = \{V_1, \dots, V_n, V_1', \dots, V_n'\} \\ relation_j = \{(v, v') \in D^n \times D^n | I(v) = I(v')\} \end{gathered} \tag{4}$$

Another related concept under the umbrella of the invariant principle is the monovariants, also called semi-invariants. These describe quantities that change consistently in only one direction, either increasing or decreasing [35].

In general, Invariant constraints are commonly found in:

- Chemical systems, where total mass, energy, or atomic counts are preserved.
- Mechanical systems, with conserved momentum or angular momentum.
- Networked systems, where flow conservation must be applied at nodes.

When encoded in a CN, invariants can drastically reduce the search space by excluding any solution path that violates these fundamental, unchanging properties of the system. Their presence enforces physical plausibility and logical consistency at a structural level.

These constraint classes, and particularly the role of invariants, provide a structural scaffolding for building CN-based models of real-world systems. When the constraints arise from physical laws, such as conservation principles, rate equations, or topological rules, the CN becomes physics-informed, an idea that will be developed further in the next section.

2.3. Physics-Informed Models

When modeling real-world systems or processes, it becomes inefficient, and often misleading, to rely purely on abstract or data-driven models. Physical systems or processes are not arbitrary. Their admissible states must be not only logically consistent, but also compliant with conservation laws, thermodynamic limits, constitutive relationships, geometric/topological compatibility, and other domain-specific principles. Embedding such prior knowledge into a mathematical model provides a foundation for incorporating physical realism into computational and data analysis. This motivates the development of physics-informed formulations grounded in the structure of CNs, which constitute the basis for the models proposed in this work.

Physics-informed models have emerged as a powerful approach for integrating physical laws directly into mathematical or computational formulations, particularly in data-driven contexts. They are most commonly associated with machine learning frameworks, such as physics-informed neural networks (PINNs), in which physical laws are embedded as differential constraints or regularization terms in the loss function [11]. In this work, a broader interpretation of the concept is adopted and applied within the framework of CNs. This leads to the formulation of what is hereby referred to as a Physics-Informed Constraint Network (PICN).

A PICN refers here to any CN instance in which domain knowledge, expressed through physical laws or system's constituent relationships, is embedded directly into the network's structure. This is achieved by incorporating observable variables of the system into the variable set, and encoding governing principles (e.g., conservation laws, stoichiometry, or structural compatibilities) as hard or soft constraints. These constraints act as structural priors that reduce the solution space to configurations that are physically meaningful. The result is a PICN model capable of encoding system behavior while preserving physical consistency, potentially enhancing its interpretability, numerical stability, and predictive performance.

In formal terms, let the system be defined over a state space D^n, with an observable vector $v \in D^n$, and let $L_{phys}(v)$ denote a set of physical constraints derived from known laws (e.g., balances, kinetics, geometry). In general, a physics-informed CN goal can be defined descriptively as:

$$Find\ v \in D^n\ such\ that \begin{cases} C_{logic}(v) = true & (logical\ constraints) \\ L_{phy}\ (v) = constant & (physical\ invariants\ /\ laws) \\ f(v)\ minimized\ (optional) & (objective\ function\ if\ any) \end{cases} \quad (5)$$

Where:

- C_{logic} is a set of logical or structural constraints (e.g., variable bounds, dependencies),
- L_{phys} includes constraints such as:
  - Mass balance: $\sum_i m_i^{in} = \sum_j m_j^{out}$
  - Elemental conservation (e.g., carbon): $\sum_k C_k * n_k = C_{total}$
  - Energy conservation: $H_{in} = H_{out}$
- $f(x)$ is a target objective, such as minimizing estimation error or maximizing data consistency.

Physics-informed formulations have demonstrated utility in different fields such as computational fluid dynamics, constraint programming, chemistry, chaotic systems, among others [36]. By embedding physical laws into the problem's structure, they enhance numerical stability in simulations and extrapolations, constrain results to remain physically plausible, and reduce reliance on large datasets particularly in cases where purely data-driven models can be prone to overfitting, instability, or non-physical predictions.

In the context of application of the present work, which is thermochemical recycling of plastics via steam cracking, the measurable outputs (e.g., species yields, carbon balances) are obtained from a conversion process governed by physical and chemical laws which define invariant relationships among variables. Physics-informed formulations are thus used to leverage this structure and develop models towards:

- Detect and correct inconsistencies in measured data,
- Predict unmeasured species (interpolation or extrapolation),
- Reveal feedstock-to-product relationships through constrained parameter estimation.

The details of the framework's implementation will be presented in Section 4.

3. Statistical Systems and Radicals Formation Overview in Steam Cracking Processes

3.1. Stochastic Processes and Statistical Systems: Definition and Overview

The study of phenomena with inherent randomness events has given rise to the concepts of stochastic processes and statistical systems, two interconnected fields that provide useful conceptual frameworks for understanding and modeling complex behaviors across diverse disciplines. A stochastic process is a collection of random variables, representing events of a system, indexed by an ordered parameter (e.g., time or space). In other terms, it can be seen as a mathematical construct used to model systems that evolve over time or space under the influence of inherent randomness. In contrast to deterministic systems, where the future state is predictable and follows a defined behavior given initial or present conditions, in stochastic systems the outcome of a process is not uniquely determined, even with complete knowledge of the current state, but instead governed by probability distributions.

A statistical system emerges when a stochastic process reaches a stationary regime or when the ensemble of its possible outcomes can be described by consistent statistical measures [37]. In such systems, the properties of interest (e.g., mean, variance, correlation structure) are not associated with single trajectories, but with distributions over an ensemble. In physical systems, especially those involving random fragmentation or molecular interaction dynamics (as in the thermochemical conversion process described in this work), the outcome of such fragmentations when a steady state has been achieved can be represented by a probability distribution over the domain of possible product states. As an example, a statistical system composed of molecular fragments can be characterized by a discrete distribution:

$$P(k) = P(X = k), k \in N \tag{6}$$

where X is the random variable representing a particular outcome that summarizes the stochastic process in statistically steady state (e.g., carbon number, chain length of the product species, etc.), and $P(k)$ denotes the probability of observing a fragment with some characteristic k within the X domain.

For such systems, a probability mass function (PMF) in the discrete case or a probability density function (PDF) in the continuous case can be defined. These functions may vary according to parameter sets associated with external conditions, providing a generalized representation of the system's behavior. These distributions must satisfy:

$$\sum_{k \in K} P(k, \theta) = 1 \ (discrete) \ or \ \int_X p(x, \theta) dx = 1 \ (continuos) \tag{7}$$

Where $K \subseteq \mathbb{N}$, or $X \subseteq \mathbb{R}$, and θ correspond to a set of parameters which define the function's topology.

This approach also enables the incorporation of additional modelling constraints derived from such intrinsic features of the statistical system, which may carry structural or physical meaning at a higher level of abstraction. Such a framework provides a useful basis for generalizing real-world random processes outcomes, such as the free-radicals phenomena occurring in the thermochemical conversion of plastics, which will be explored in subsequent sections.

3.2. Energy Bond Dissociation and Free-Radical Reaction Mechanisms in Steam Cracking

The thermal cracking of carbonaceous feedstocks, such as plastics and hydrocarbons, is fundamentally governed by free-radical chemistry. The aim is to break down large hydrocarbon molecules into smaller and more valuable species to be used as raw materials in chemical synthesis. These reactions, driven by thermal energy, begin with the homolytic cleavage of chemical bonds which is a process where a covalent bond breaks symmetrically, generating two free radical species, each with an unpaired electron making them highly reactive.

At the elevated temperatures of steam cracking, the kinetic energy of molecules becomes sufficiently high to overcome the activation energy barriers for the bond's cleavage. Such energy requirement is defined by a scalar in kJ/mol known as the bond dissociation energy (BDE), a key thermodynamic parameter that correlates with the likelihood and selectivity of the radical initiation. In general, weaker bonds require less energy to cleave and are thus more prone to scission at lower temperatures or shorter residence times. In homogeneous structures, BDE may serve as a primary indication of the stability of the resultant radical, with low BDE indicating a higher radical stability, but in general cases, other aspects as the radical structure should also be considered [38,39].

The BDE is generally determined by the bond type, for instance, C-C bonds generally have BDEs ranging from 330 to 370 kJ/mol, while C-H bonds are stronger, typically ranging from 410 to 440 kJ/mol [40]. However, BDE is also strongly influenced by the local chemical environment within the molecule. Structural features such as branching, conjugation, aromaticity, or the presence of electron-withdrawing or electron-donating groups can alter the electron density around a bond, thereby modifying its stability and energy of cleavage. For example, bonds of carbons adjacent to an olefin (allylic carbon) or aromatic rings (benzylic) often exhibit lower BDEs and are more prone to scission (e.g., allylic C–H in propene ~370 kJ/mol; benzylic C–H in toluene ~377 kJ/mol [40]). This is due to delocalization of radical electrons into the π-system, which stabilizes the resulting radical species by resonance. In contrast, bonds embedded in rigid cyclic structures (e.g. bycycloalkanes) or alpha to strongly inductive electron-withdrawing groups (where the resulting radical is destabilized) may require more energy to break (e.g., C–C in CF_3–CH_3 ~410 kJ/mol).

Overall, the more homogeneous the bond dissociation energy is along the chain, the more random the occurrence of the cleavage process. In heterogeneous polymeric chains, the electronic cloud distortion lead by certain molecular motifs serves as preferential cleavage sites along the chain. This structural dependency creates asymmetries in the radical initiation step, and thus influences the initial radical pool composition, ultimately shaping the pathways and outcomes of the secondary reaction network in steam cracking processes.

The initial bond dissociations, often known as primary free radical reactions or initiation steps, are crucial for generating the first radicals that propagate the chain [41,42]. Once initiated, these primary free-radical reactions describe the immediate decomposition of the parent molecule, where scissions occur along the chain to produce radical intermediates. As discussed earlier, these reactions will be largely driven by the statistical distribution of BDEs within the molecular structure and the thermal environment, and they will occur in the condensed or early gas phase. In general, the rate of initiation is highly temperature-dependent, following an Arrhenius-type relationship ($\sim e^{-a/T}$).

The produced radicals are chemically reactive and short-lived (nano to microseconds), and their formation represents the onset of a dynamic and probabilistic chain of transformations [43,44]. For a generic alkane, an initiation step might involve the homolytic cleavage of a C-C bond:

$$R - R' \xrightarrow{\Delta} R \bullet + R' \bullet$$

where $R \bullet$ and $R' \bullet$ are free radicals, such as alkyl radicals, with an unpaired electron. These radicals, abstract hydrogen atoms from other molecules, producing new radicals and stable molecules (e.g., $R \bullet + R'H \rightarrow RH + R' \bullet$). This propagation step sustains the chain reaction. Additionally, radicals may undergo β-scission, where a C–C bond adjacent to the radical center breaks, forming an olefin and a smaller radical (e.g., $RCH2CH2 \bullet \rightarrow R \bullet + CH2 = CH2$). This path is critical in steam cracking, as it generates valuable unsaturated products like ethylene and propylene, and it is in general favored at high-temperature environments.

Following the initial bond scissions, a radicals' pool is formed where the system enters a regime of secondary reactions involving the further transformation of primary products in the gas phase and determine the final molecular composition of the product gas. Radicals can react with olefins or with other radicals, leading to addition reactions that form larger radicals (e.g., $R \bullet + CH2 = CH2 \rightarrow RCH2CH2 \bullet$), which later may undergo further scission or termination. Termination occurs when two radicals combine to form a condensed stable molecule (e.g., $R \bullet + R' \bullet \rightarrow R - R'$) or disproportionate, producing an alkane and an olefin. Secondary reactions also include radical isomerization (typically involving intra or intermolecular hydrogen atom transfer), and can lead to a variety of outcomes, alkene isomerization and cycloisomerization, which influence further the product distribution [43,44].

Under certain conditions, intramolecular cyclization occurs along with specific inter radical recombinations and olefinic reactions (such as Diels-Alder), leading to the formation of aromatic species such as benzene, toluene, xylenes, and styrene (BTXS). These reactions become significant especially in feedstocks rich in olefins or polymers containing aromatic rings. Additionally, as radical intermediates continue to react, aromatic growth mechanisms may occur, such as hydrogen abstraction-acetylene addition (HACA), enabling polycyclic aromatic hydrocarbon (PAH) formation [45,46]. These larger aromatic compounds can eventually condense into soot precursors, particularly in localized reaction zones with elevated temperature or insufficient hydrogen availability [47,48]. The formation of such species is in general a concern as they affect product quality and, in certain systems, can lead to reactor fouling or coke deposition. Thus, the interplay of these secondary pathways and their control is essential not only for the optimization of the light product distribution but also to regulate the potential formation of such undesired by-products.

3.3. Applicability of Statistical-Based Data Analysis Modeling to Thermochemical Conversion Systems

Overall, the transformation from feedstock to final product via free-radical chemistry is inherently stochastic. As presented before, each radical pathway represents one among many possible sequences of elementary reactions, forming an intricate network of random occurring events at molecular scale. Nonetheless, from the collective result of a vast number of parallel and competing radical molecules and pathways, emerge a fully formed set of stable species that constitute the final observed product species distribution.

Notably, the more the system is allowed to explore a broader range of accessible microstates (i.e., limited selectivity), the more the product spectrum will reflect the intrinsic randomness of the process. Consider the polymeric breaking down process: since the rate of initiation can be taken as a function of temperature and the BDE distribution, the random behavior will become most apparent as the reactor's temperature goes up and with feedstocks containing structurally homogeneous hydrocarbons, such as linear alkanes or polyolefins. The pyrolytic reaction progresses as the resulting molecules try to adopt more stable structures and lengths. As a rule of thumb, the shorter the chain, the more stable it becomes. Then, as mentioned in Section 3.1, whenever the outcome of such stochastic decomposition leads to repeatable patterns that can be arranged in an ordered sequence form (e.g. carbon number), the patterns recurrence can be summarized into a discrete probability mass function defined by a set of parameters (See Equation (6) and Figure 3).

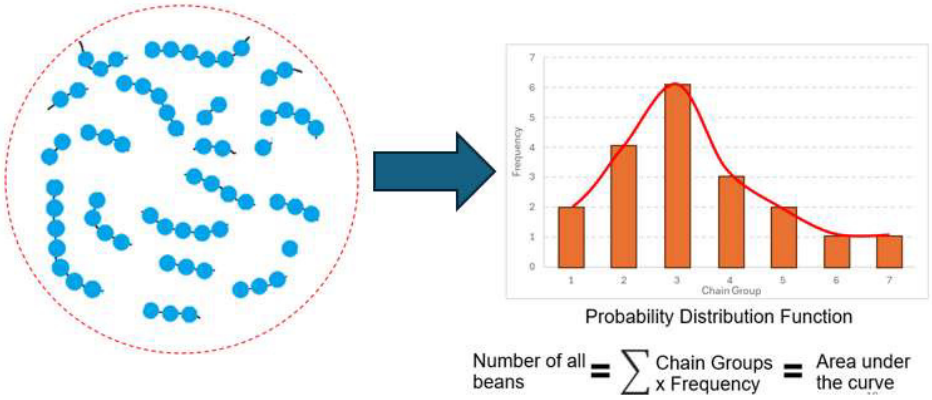


Figure 3. Conceptual schematic of how a distribution function can be constructed by counting the different chain species produced from the radical system according to their carbon group k.

So, given the absence of external selectivity controls (e,g. a selective catalyst), the inherent stochasticity of the linear radical chemistry, where initiation, propagation, and termination events occur with random spatial and temporal characteristics, will give rise to emergent product distributions that can be well-described by statistical models.

To illustrate this point, consider the next example. In its simplest form, the set of events in a homogeneous polymeric chain within a hot environment can be conceptualized as a system that comprises two types of elements: broken and unbroken bonds. If N_0 is the number of initial bonds, the probability q of finding a bond in the chemical system is defined by $q = N_b/N_0$, with N_b being the number of remaining bonds at time t. Then, the probability to find a broken bond is $p = 1 - q$ (as this can also be taken as the probability of breaking a bond at time t, it is a parameter that is directly related to the temperature in a pure thermally driven system). Assuming independence of events, the probability of finding n consecutive unbroken bonds in n number of blinded pickups in the system is defined by q^n. A chain molecule of k carbons consists of a set of $k - 1$ unbroken bonds and 2 broken bonds at its extremes. Therefore, the probability of finding such a set of elements in the system will be p^2q^{k-1}. From the carbon´s perspective, the probability of finding a particular carbon in such a set will be $P(k, p) = kp^2q^{k-1} = kp^2(1-p)^{k-1}$. This is commonly known as the Flory-Schultz distribution; it describes the probability of obtaining a chain of length k after the occurrence of random and identically independent scission events with probability p (see Figure 4). It should be noted that this is a special case of the well-known Negative Binomial Distribution for $k - 1$ successes and exactly 2 failures.

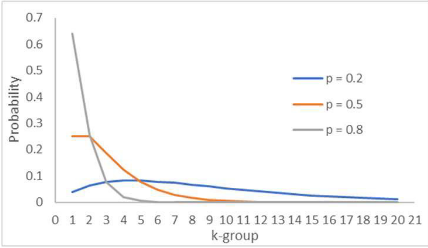


Figure 4. Flory-Schultz distribution.

This simple conceptualization illustrates the increasing abundance of shorter chains as the probability of a breaking event increases. Although the real-life cases may be more complex, this simplified view offers a clear picture of the generalization that the distribution function brings to the results obtained from the stochastic process. Such mathematical abstraction also facilitates the integration of additional structural constraints, into the modeling framework. This perspective provides the conceptual foundation for the data analysis modelling applied over the outcomes of the thermochemical process evaluated in this work, and sets the basis for the parametric system model developed in **Paper I** which will be outlined in Section 4.1.

On the other hand, molecular processes like steam reforming or gasification reactions do not exhibit such stochastic fragments-like behavior. These systems are better characterized as deterministic systems, where the product formation is governed by well-defined reaction pathways, clear rate-determining steps, specific stoichiometries, and rate laws. The outcomes of these reactions are not the result of a large ensemble of probabilistic paths from stochastic fragmentation, but of controlled molecular interactions mediated by different reactants or catalytic surfaces, leading to predictable formation of small molecules such as H_2, CO and CO_2. As such, their products cannot be meaningfully described by a distribution function over discrete molecular configurations. Instead, they can be modeled using mechanistic or equilibrium-based kinetic models based on reaction rates, surface coverage, and/or equilibrium thermodynamics.

Now, considering heterogenous feedstock systems, the final product species distribution will depend heavily on the polymeric composition and reaction environment. For example, in mixtures containing aromatic polymers (e.g., polystyrene or PET) or precursors with strong functional groups (e.g., halogens, oxygenates), the radical pathways become biased due to the presence of specific reactive characteristics. These motifs alter bond dissociation patterns, introduce new reaction channels (such as cyclization and aromatic growth), and lead to selective production or suppression of some species more than others. Similarly, the use of selective catalysts, if introduced, may break the stochastic symmetry of the process by enhancing or suppressing specific reaction pathways of certain species, thus affecting the randomness of the process and skewing the distribution.

In that case, the evaluation of the product distribution requires a new modeling approach. It becomes necessary to consider the structural characteristics of the feedstock, including bond types, functional groups, and connectivity. Such characteristics influence the reactivity landscape and must be encoded explicitly into a modeling framework that can integrate a feature-level of abstraction. This motivated the development of the carbon bond-based system model, which incorporates relevant chemical structure with bond-type level attributes to evaluate correlations between product distributions and heterogeneous feedstock systems. Such an approach opens the way for data analysis modelling that generalizes the complexities of the radical polymer decomposition driven by heterogenous feedstocks but still preserves key information of its chemistry. This is the foundation of the modeling strategy followed in **Paper II and VI** and described in Section 4.2.

4. Integration of Physics-Informed CN Frameworks with Data Analysis Models

As described in Section 2.1, CNs provide a structured way to represent real-world phenomena by framing them in terms of variables, domains, and constraints that can be systematically explored. Their usefulness lies in the incorporation of different constraining layers to delimit the space of admissible outcomes. When applied to thermochemical conversion processes such as steam cracking, these frameworks offer the possibility of enclosing the complexity of feedstock–product transformations within a mathematically tractable form. In this way, the derived hybrid models can move beyond black-box data fittings and become meaningful representations anchored to the physical principles governing the phenomena, making them valuable tools for both outcome evaluation and prediction.

Building on the system's data outcomes, the first structural layer that can be implemented for a thermochemical system arises from the conservation laws. These introduce invariants, such as the elemental balances of carbon, hydrogen, and oxygen which must hold across the system's transformation regardless of the complexity of the reaction network. Their role is comparable to that of state functions: they depend only on the total amounts entering and leaving the system, not on the particular sequence of transformations in between. By enclosing the entire feed-to-product conversion within these invariants, a fundamental frame can be established in which the redistributed product composition remains fully constrained by the same conserved totals (Figure 5 shows the concept's schematic for a closed system thermochemical conversion process).

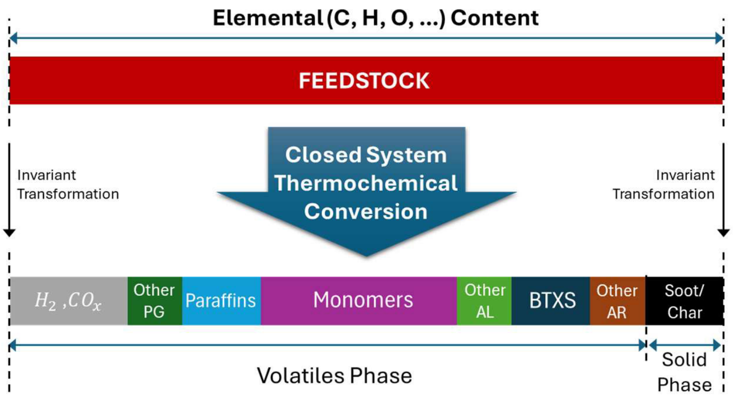


Figure 5. Schematic of main invariant enclosure in a closed system transformation (PG: Permanent gases, AL: Aliphatics, BTXS: Benzene + Toluene + Xylenes + Styrene, AR: Aromatics)

After the main invariants, which provide the broadest enclosure of the system, additional constraint layers can be unlocked by going deeper into the governing principles of the process. These layers restrict the solution space to reflect more tightly the physical and chemical characteristics of the evaluated system. Examples include statistical regularities in chain scission, structural limitations imposed by bond hybridization and valence, and the specific reactivity of different carbon environments.

A central challenge in modeling the system under study lies in its high dimensionality. Large numbers of species and complex feedstock mixtures quickly overwhelm conventional data-analysis techniques. Dimensionality reduction is therefore essential. Purely data-driven techniques, such as

Principal Component Analysis (PCA), attempt to address this by compressing the data based on certain statistical metrics (e.g. the variance) [49]. However, such approaches often disregard physical admissibility, leading to representations that may violate conservation laws or lack physical interpretability. In contrast, the approaches developed here pursue dimensionality reduction through the construction of reduced PICN instances, ensuring that the resulting models capture the essential variability of the data while remaining compact, consistent with conservation principles, and physically meaningful.

The methodology of progressive layering of constraints and dimensionality reduction is reflected in the two models presented in the following subsections. In Section 4.1, the Parametric System Model (PSM) focuses on the product side, using probability distributions to capture regularities in species yields and to extrapolate unmeasured product groups. In Section 4.2 and 4.3, the Carbon Bond Group (CBG) model maps both feedstock and products into a chemical structural features framework, enabling the study of correlations between input composition and output distributions. Together, these approaches illustrate how key chemical insights can be translated into structural constraints to create low-dimensional CN-based data analysis models.

Figure 6 presents an overview of the general methodology used to structure and deploy the PICN instances developed in this work.

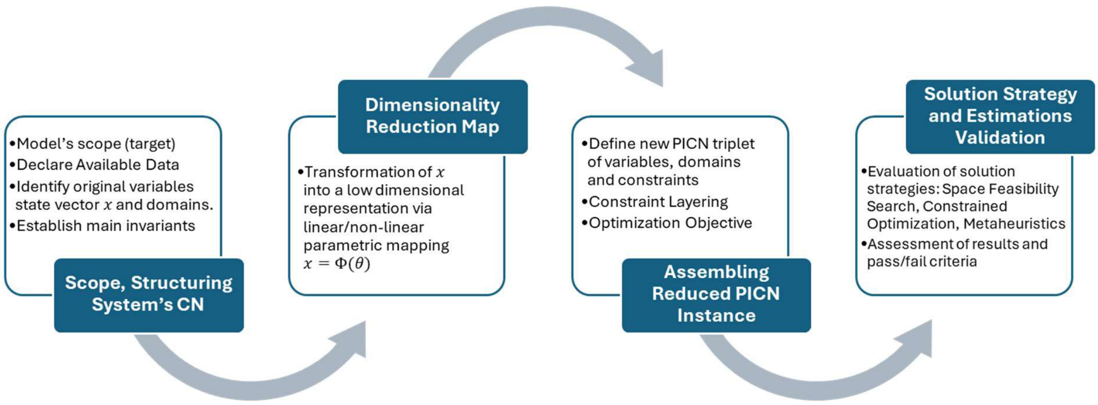


Figure 6. General methodology of setting up the PICN-based models in this work.

4.1. Parametric system model development for data quality analysis

In any thermochemical conversion process, the outcomes of experimental characterization are subject to three main sources of variation: changes to the feedstock, alterations to the operational thermodynamics, and measurement-related errors or limitations in the setup. Therefore, a data analysis model intended to correlate and evaluate the acquired data in relation to such sources must be inherently linked to the chemical nature of the process.

In line with this premise, the approach followed in **Paper I** involves the processing of the species yields data from steam cracking conversion using a mathematical transformation referred to as the Parametric System Model (PSM) (Figure 7). The defining characteristic of this model is that the high-dimensional system representing the original data of product species, is reformulated into a lower-dimensional PICN instance tractable to solve through parametric distribution functions with well-defined topologies. These functions act simultaneously as dimensionality-reduction tools and as carriers of physical and statistical constraints. The PSM's goal is to create a highly constrained system that be tightly fitted to the nature of the conversion process. By doing so, the model is intended to provide physically consistent outcomes and capable of: (i) identifying inconsistencies in the measured data, (ii) estimating unmeasured quantities or species that may be inaccessible due to equipment limitations, and (iii) evaluating certain measurements relevance for the model's outcomes, so as to minimize the sampling characterization effort.

In Paper I, the focus is on the application of the model for prediction of unmeasured species and data quality assessment of the data coming from steam cracking of a polyolefinic feedstock (polyethylene). However, the concept of creating such a PICN framework based on probabilistic functions constraining the data can also be applied to other types of feedstocks and mixtures.

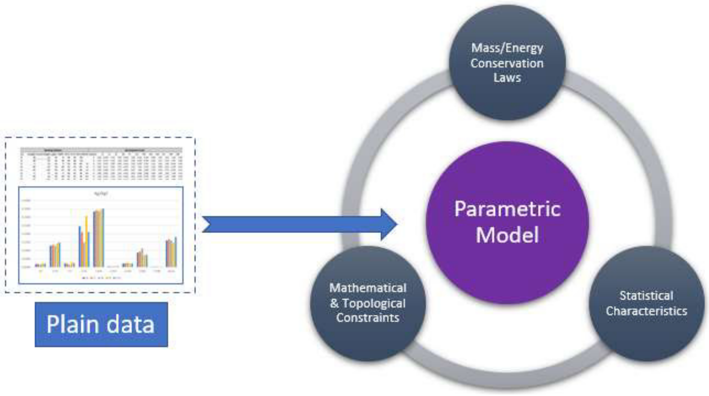


Figure 7. Conceptual schema for the "plain data" transformation into the proposed parametric model.

4.1.1. PICN Formulation and Transformation

The starting point of the PSM is its formulation as a CN. The product mixture is classified into distinct molecular systems: paraffins, olefins, aromatics, syngas, and char. The species of each system are described as variables $\in \mathbb{R}^+$ corresponding to their molar yields.

The model's core constraints are defined based on the system's conservation laws, which define the main invariants of this original CN. In **Paper I**, only mass conservation is taken into account and expressed in terms of the main elemental balances. Therefore, for a feedstock that is composed of carbon (C), hydrogen (H), and oxygen (O), the balances are mathematically defined using molar fractions and expressed according to Equations (8)–(10).

$$rel1 = \left\{ n_{C,tot}^{pf} \sum_{k=1}^{\infty} X_{C,k}^{pf} + n_{C,tot}^{of} \sum_{k=1}^{\infty} X_{C,k}^{of} + n_C^{arom} + n_C^{CO} + n_C^{CO_2} + n_C^{char,out} = n_C^{fuel} + n_C^{char,in} \right\} \tag{8}$$

$$rel2 = \left\{ n_{C,tot}^{pf} \sum_{k=1}^{\infty} X_{C,k}^{pf} \gamma_k^{pf} + n_{C,tot}^{of} \sum_{k=1}^{\infty} X_{C,k}^{of} \gamma_{k,eff}^{of} + n_{H,tot}^{arom} + n_H^{H_2,out} + n_H^{H_2O,out} = n_H^{fuel} + n_H^{H_2O,in} + n_H^{H_2,in} \right\} \tag{9}$$

$$rel3 = \left\{ n_O^{CO} + n_O^{CO_2} + n_O^{HC,out} + n_O^{H_2O,out} + n_O^{O_{ext},out} = n_O^{fuel} + n_O^{H_2O,in} + n_O^{O_{ext},in} \right\} \tag{10}$$

$$\gamma_k^{pf} = \left(2 + \frac{2}{k}\right); \quad \gamma_{k,eff}^{of} = \sum_{g=0}^{2\lfloor k/2 \rfloor - 1} X_{C,k,g}^{of} \left(2 - \frac{2g}{k}\right) \tag{11}$$

Here, n_a^s and $X_{a,k}^s$ correspond, respectively, to the number of moles and the molar fraction of the element $a \in \{C, H, O\}$ and the system $s \in \{pf, of, arom, \dots\}$, with pf, of and $arom$ indicating the paraffin, olefin and aromatic groups, respectively. The subindex k represents the number of carbons in the paraffin or olefin chain and referred to as the carbon group (or the k-group). tot indicates the total amount. The model's implementation was done over paraffins and olefins systems, as these are linear systems that emerge from the chain scission events that follow repeatable, size-ordered patterns (see Section 3.3). Their carbon-number distributions can therefore be expressed as ordered sequences, well suited for probabilistic modeling, and is also commonly used in the petrochemical industry to compare yields of specific species sizes and to assess the quality of products to be used as liquid fuels [50,51]. The model's concept could be extended to the aromatics molecular system through some geometrical arrangement, based on their rings or on their primordial radicals, but this is out of the scope of the present work.

In Equation (9), the moles of carbon and hydrogen can be interconverted in hydrocarbon chains by the means of γ_k^s, which refers to the molar hydrogen to carbon (H/C) ratio at a chain k of the system $s \in \{pf, of\}$. $\gamma_{k,eff}^{of}$ corresponds to the olefins effective molar H/C at a particular k and is defined as the weighted molar fraction sum of the H/C ratios for all possible olefin species present in k (see Equation (11)). The subindex g corresponds to the olefin species group defined by the number of unsaturated bonds minus one of the chain k, starting from the mono-ene case at $g = 0$ ($\lfloor \cdot \rfloor$ indicates the floor integer function).

In Equation (10), $n_O^{HC,out}$ corresponds to the moles of oxygen in oxygen-containing hydrocarbon species (e.g., some oxygenated aromatic species). The term $n_C^{char_{in}}$ accounts for external sources of char/soot entering the system, for instance, char transported from the combustor side during the DFB cycle. In the model's estimations, contributions of carbon to the product slate from char/soot of external sources are assumed negligible in the considered conversion process at the cracker reactor.

Relative to the hundreds of thousands of monomer units present in a real polymeric feedstock, extending the series to the infinite is essentially a mathematical formality to avoid restrictions imposed on the length of the species that can be formed in the cracking process. Moreover, these series must be convergent, which imposes a mathematical constraint on the sequence behavior and the transformations associated with them.

Up until now, the direct representation of the product mixture as individual species conforms in essence a high-dimensional CN. Since each species corresponds to a variable, the full system rapidly becomes intractable due to the large number of possible molecules and the limited coverage of the experimental setup. Hence, in the proposed PSM model a transformation is introduced in which the groups of chemically related species are described by parametric distribution functions. This mapping reduces the effective dimensionality of the original network, augmenting it with topological constraints while preserving its essential chemical characteristics.

Consider the molar fractions $X_{a,k}^S$, with $a \in \{C, H\}$ and $s \in \{pf, of\}$. Rather than treating each fraction as an independent variable, the PSM represents them as outcomes of a parametric family defined as Equation (12).

$$X_{a,k}^s = f_a^{pf}(k; \boldsymbol{\alpha}_a^s) \tag{12}$$

Where f_a^S is a discrete function with a semi-infinite support $k \in \mathbb{N}^+$, ruled by a low-dimensional parameter vector $\boldsymbol{\alpha}_a^s$, defined for each system s and element a. This corresponds with the model's key mathematical transformation.

Through this formulation, the PSM creates a compact representation in which the original large sequence of potential decision variables of the product mixture is projected onto a small number of parameters, thereby compressing the original high-dimensional CN into a reduced PICN. The advantage of this approach is that it can provide a closed-form solution to the equation system formed by Equations (8) and (9) and, at the same time, delivering the necessary topology and convergence criteria needed to satisfy the model's mathematical constraints.

4.1.2. Model Implementation and Additional Constraints

Following the parametric transformation of the paraffin and olefin sequences, the objective is to define a function f_k that can fit the measured data using the minimum number of shape parameters while remaining mathematically consistent. In general, to ensure physical and mathematical consistency, the function's formulation must satisfy a set of constraints:

1) Convergence: the parametric functions must be positive and yield sequences that decay with k and converge as $k \rightarrow \infty$, preventing unrealistic accumulation of mass at high carbon numbers. The function's form and predicted area must adhere to the imposed conservation laws.
2) Topology: the functions are restricted to monotonic or unimodal positively skewed forms, reflecting the fragmentation patterns of polymer chains in steam cracking (See Section 3.3).
3) Flexibility: it must be flexible enough to handle relatively large changes in species concentration, while remaining convergent and aligned with the measured data
4) Dimensionality: it needs to be defined with the fewest number of parameters to avoid overfitting. In this way, the model attains the dual goal of compressing the system's dimensionality while retaining the essential physical information contained in the measurements and other constraints.

As a brief illustrative example: Let's consider a high-temperature decomposition scenario following the Flory-Schultz distribution with p=0.7 shown in Figure 4. Suppose the experimental setup yields data only for the first three aliphatic carbon group species. In the absence of prior information regarding the system's decomposition, the yields decreasing trend of the measured species serve as the initial indicator of the system's behavior.

For this example, considering the mass conservation law, approximately 90% of the normalized mass in the distribution will lie on the measured side. This significantly limits the range of viable estimations for $k \geq 4$. The topology must reflect such mass and the behavior of measured data, while also ensuring convergence in the limit to infinity. At this stage, the system becomes highly constrained. Assigning a solution function with an excessively large, insufficient, or non-convergent area of the unknown side should be ruled out (see Figure 8). Alternatively, if the measured area is excessively long or short so that it hinders a convergent fit within boundaries, it is likely the experimental data must be reviewed and check for potential data quality or experimental setup issues.

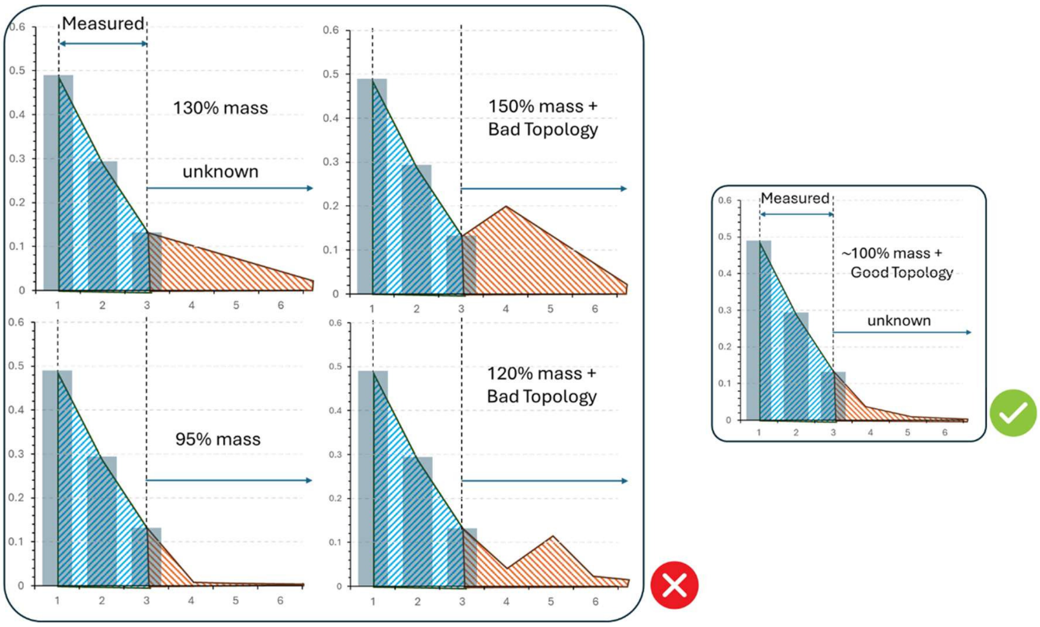


Figure 8. Examples of incorrect and correct cases of function's fitting.

In terms of the type of functions, the realm of discrete distribution functions can provide the candidates that best fulfill all of the aforementioned conditions. Several candidate functions are available, ranging from mono-parametric to bi-parametric forms, such as geometric, Flory–Schulz, Poisson, or negative binomial distributions, as well as discretized continuous functions with heavier tails [52]. This forms the set of this categorical variable. Table 1 presents the various probability distribution functions evaluated in **Paper I** study.

Table 1. List of mono-parametric and bi-parametric distribution functions used in Paper I. The asterisk-marked rows correspond to heavy-tailed distributions. $i = 1$ correspond to $k = 1$ for paraffins and $k = 2$ for olefins (Note that Γ is the mathematical gamma function, γ_{inc} and Γ_{inc} are the lower and upper incomplete gamma function respectively, and Weibull is heavy-tailed for $0 \leq \alpha_2 \leq 1$).

Distribution	**Mathematical Expression**
Geometric	$f_i(\alpha) = \alpha(1-\alpha)^{i-1};\ \ 0 < \alpha < 1$
Flory-Schultz	$f_i(\alpha) = \alpha^2 i(1-\alpha)^{i-1};\ \ 0 < \alpha < 1$
Poisson	$f_i(\alpha) = \alpha^{i-1}e^{-\alpha}/(i-1)!\,;\ \ \alpha > 0$
Negative Binomial	$f_i(\alpha_1,\alpha_2) = \Gamma(i+\alpha_2-1)\alpha_1^{\alpha_2}(1-\alpha_1)^{i-1}/\big(\Gamma(i)\Gamma(\alpha_2)\big);\ \ \alpha_2 > 0, \alpha_1 \in [0,1]$
Conway-Maxwell	$f_i(\alpha_1,\alpha_2) = \alpha_1^{(i-1)}/\big(\Sigma_{j=0}^{\infty}(\alpha_1^j/(j!)^{\alpha_2})((i-1)!)^{\alpha_2}\big);\ \ \alpha_1,\alpha_2 > 0$
Burr *	$f_i(\alpha_1,\alpha_2) = (1+(i-1)^{\alpha_2})^{-\alpha_1} - (1+i^{\alpha_2})^{-\alpha_1};\ \ \alpha_1,\alpha_2 > 0$
Fréchet *	$f_i(\alpha_1,\alpha_2) = e^{-(i/\alpha_2)^{-\alpha_1}} - e^{-((i-1)/\alpha_2)^{-\alpha_1}};\ \ \alpha_1,\alpha_2 > 0$
Dagum *	$f_i(\alpha_1,\alpha_2) = (1+i^{-\alpha_2})^{-\alpha_1} - (1+(i-1)^{-\alpha_2})^{-\alpha_1};\ \ \alpha_1,\alpha_2 > 0$
Gompertz	$f_i(\alpha_1,\alpha_2) = e^{-\alpha_1(e^{\alpha_2(i-1)}-1)} - e^{-\alpha_1(e^{\alpha_2 i}-1)};\ \ \alpha_1,\alpha_2 > 0$
Weibull *	$f_i(\alpha_1,\alpha_2) = e^{-((i-1)/\alpha_1)^{\alpha_2}} - e^{-(i/\alpha_1)^{\alpha_2}};\ \ \alpha_1,\alpha_2 > 0$
Gamma	$f_i(\alpha_1,\alpha_2) = \gamma_{inc}(\alpha_1,\alpha_2 i)/\Gamma(\alpha_1) - \gamma_{inc}(\alpha_1,\alpha_2(i-1))/\Gamma(\alpha_1);\ \ \alpha_1,\alpha_2 > 0$
Lomax *	$f_i(\alpha_1,\alpha_2) = (1+(i-1)/\alpha_1)^{-\alpha_2} - \big(1+i/\alpha_1\big)^{-\alpha_2};\ \ \alpha_1,\alpha_2 > 0$
Gamma Inverted *	$f_i(\alpha_1,\alpha_2) = \Gamma_{inc}(\alpha_1,\alpha_2/i)/\Gamma(\alpha_1) - \Gamma_{inc}(\alpha_1,\alpha_2/(i-1))/\Gamma(\alpha_1);\ \alpha_1,\alpha_2 > 0$

In general, the flexibility level of a function increases with the number of parameters that it incorporates. In particular, the family of bi-parametric functions are highly adaptable for describing diverse phenomena across various knowledge fields. Among these functions, heavy-tailed distributions (asterisks in Table 1) are especially robust against large variations within a dataset. This flexibility arises from their characteristic tails, which decay at a slower rate than functions with exponential decay; making them well-suited for modeling extreme events, such as the ones found in natural phenomena, e.g., survival times, river discharges, etc. [53].

For the hydrogen function case, an additional constraint can be formulated to the behavior of the H/C ratio. For paraffins, hydrogen content is directly derived from valence rules ($2k + 2$, see Equation (11)), while for olefins the effective hydrogen ratio is restricted by the admissible range of unsaturation states. The upper limit is set by the mono-ene case given by Equation (13), while the lower limit corresponds to the fully unsaturated scenario, expressed in Equation (14). An intermediate case is used as reference and defined at the fully conjugated state (intercalated double bonds in the chain) as per Equation (15). The olefin´s hydrogen function must fall within this reduced solution space to ensure that the fitted distributions are not only statistically adequate but also chemically admissible.

$$\gamma_{k,high}^{of} = 2n_{C,k}^{of} \tag{13}$$

$$\gamma_{k,low}^{of} = \frac{1}{k}((2k+2) - 4\lfloor k/2 \rfloor)n_{C,k}^{of} \tag{14}$$

$$\gamma_{k,mid}^{of} = \frac{1}{k}((2k+2) - 2\lfloor k/2 \rfloor)n_{C,k}^{of} \tag{15}$$

In essence, the fundamental approach followed in **Paper I** involves utilizing the measured data as anchor points with its experimental uncertainties as space constraints, and Equations (8)-(9), to obtain the set of equations to estimate the specific shape parameters α_a for a given function in Equation (12). Together, these constraints define a reduced but physically consistent solution space that can be explored by optimization methods. If a finite and real set of parameters can be identified, the corresponding function becomes a solution to the PICN system. Such constraint enforcement during the parameter exploration stage guarantees to the model that the extrapolated distributions remain physically meaningful beyond the measured carbon groups.

The unique nature of most distribution functions requires the application of non-linear solvers or Monte Carlo methods to find the parameter set that fulfills all the constraints imposed. Once the shape parameters are determined, the function can be visualized by overlaying it onto a column bar graph representing species in mol/kgf versus the respective k-groups.

4.1.3. Experimental Data Validation Method

The topologic constraints and invariants can act themselves as first validation layer for the measured data (see Figure 8 and its related example). Cases in which the model fails to converge or violates balance requirements can be interpreted as signals of potential inconsistencies in the experimental data.

Nonetheless, a further validation layer can be implemented by the means of the oxygen balance given by Equation (10). This equation can be used to calculate a relevant process quantity of DFB systems known as the Bed's Oxygen Transport (BOT). This quantity can be obtained by a secondary experimental setup system and served as a benchmark to feature the model's potential for data quality assessment of the experimental results in **Paper I**.

In DFB reactors, the BOT phenomenon occurs when the bed material contains oxygen-carrying species. The circulating bed serves as a carrier medium, transferring heat from the combustor to the cracker chamber. It may also transport reactive species between the cracker and combustor. Even seemingly inert bed materials, such as silica sand, may contain traces of transition metal oxides, such as Fe_2O_3, which act as oxygen donors. Ashes carried by the bed from the combustor side can also serve as active sources of oxygen which can alter the gas product composition to some extent. Elements such as calcium participate in redox cycles [54] forming compounds such as $CaSO_4$ under oxidizing conditions within the combustor and later transformed into CaS in the reductive environment of the cracker.

By rearranging the oxygen balance in Equation (10) and applying the water H/O molar ratio of 2, the BOT can be derived as in Equation (16):

$$n_O^{CO} + n_O^{CO_2} + n_O^{HC,out} - \left(n_O^{H_2O,in} - n_O^{H_2O,out}\right) = n_O^{fuel} + \left(n_O^{O_{ext},in} - n_O^{O_{ext},out}\right) + n_O^{O_2,leak}$$

$$\Rightarrow \Delta n_O^{O_{ext}} = n_O^{O_{ext},in} - n_O^{O_{ext},out} = n_O^{CO} + n_O^{CO_2} + n_O^{HC,out} - \frac{1}{2}\Delta n_H^{H_2O} - n_O^{fuel} - n_O^{O_2,leak} \quad (16)$$

From the hydrogen balance, the level of water conversion ($\Delta n_H^{H_2O} = n_H^{H_2O,in} - n_H^{H_2O,out}$) can be estimated as:

$$\Delta n_H^{H_2O} = \left(n_{C,tot}^{pf} \sum_{k=1}^{\infty} X_{C,k}^{pf} \gamma_k^{pf} + n_{C,tot}^{of} \sum_{k=1}^{\infty} X_{C,k}^{of} \gamma_{k,eff}^{of} + n_{H,total}^{arom} + n_H^{H_2,out}\right) - \left(n_H^{fuel} + n_H^{H_2,in}\right) \quad (17)$$

The term $\Delta n_O^{O_{ext}}$ in Equation (16) describes the effective oxygen entering or leaving the reaction environment due to external sources other than the fuel or fluidization steam. Positive values indicate the external oxygen consumed by conversion products, while negative values imply oxygen removal by an external agent. In the absence of a deliberate oxygen input flow, two external oxygen sources can be identified: unintended air leakage into the cracker ($n_O^{O_2,leak}$), and the active species in the circulating bed material. Typically, leakage is minimal thanks to periodic maintenance of the reactor. However, for non-nitrogenous fuels, air leakage can be quickly calculated through the nitrogen balance with the detected level of N_2 and the air's O/N ratio. Then, Equation (17) defines the oxygen transported into the system by the bed, i.e., the BOT, and it can be estimated from the results given by the PSM implementation.

To implement this validation method, the experimental approach consisted of a High-Temperature Reactor (HTR) [55] in a parallel sampling stream (see Figure 17). In this method, gas products react further with steam at very high temperatures in the HTR (~1700°C), generating only syngas. In such case, Equation (16) takes a simplified form due to the complete reforming of the product species into H_2, CO and CO_2, as presented in Equation (18).

$$\Rightarrow \Delta n_O^{O_{ext}}\Big|_{HTR} = n_O^{CO} + n_O^{CO_2} - \frac{1}{2}\left(n_H^{H_2,out} - \left(n_H^{fuel} + n_H^{H_2,in}\right)\right) - n_O^{fuel} - n_O^{leak} \quad (18)$$

Since HTR and PSM estimations stem from the same sampled gas batch, the result from Equation (18) should align with that calculated using Equation (16) and serve as validation of the PSM estimations.

In summary, the PSM establishes a PICN that reduces the high-dimensional product space of steam cracking into a compact parametric form. By combining distribution functions with structural invariants and topological constraints, the model enables physically consistent estimation of unmeasured species, identification of measurement inconsistencies, and a systematic approach to data quality analysis at the process' downstream side. This formulation, validated against experimental datasets, demonstrates the feasibility of embedding dimensionality reduction directly into a constraint-based framework. Building on this foundation, the next section extends the concept's approach to the case of heterogeneous mixtures, where a carbon bond–based framework provides a generalized representation to capture correlations between diverse feedstock compositions and product distributions.

4.2. Heterogenous Mixtures Correlations Model: Carbon Bond Group Framework Concept

Real waste streams are rarely composed of a single polymer type. Mixed plastic fractions, textiles, and other carbonaceous residues exhibit a high degree of heterogeneity in both polymeric structure and elemental composition. Such molecular composition of a polymeric feedstock plays a crucial role in the definition of the final cracking product distribution.

4.2.1. Polymer Categories

The polymers found in plastic waste can be broadly categorized into three groups: aliphatic polymers, aromatic-containing, and heteroatom-containing polymers, each group exhibiting distinct behaviors during the thermochemical conversion process:

- Aliphatic polymers, such as polyethylene (PE) and polypropylene (PP), undergo random scission at low temperatures (>400 °C), producing fragments of varying lengths [56,57]. At higher temperatures (e.g., 700–800°C), end-chain scission occurs, favoring the formation of light hydrocarbons such as methane and ethylene, as well as aromatics and polyaromatics promoted by radical recombination and cyclization reactions [45]. Overall, polyolefins, which dominate household waste streams, typically convert almost completely into gas products at 500–600 °C. Beyond that, the gas product distribution tends towards production of monomers (C_2–C_4) and aromatics such as benzene, toluene, xylene, and styrene (BTXS), which are valuable for chemical recycling. Excessive severity, however, reduces monomer recovery and increases the risk of polyaromatic formation.

- Aromatic-containing polymers, such as polystyrene (PS) and polyethylene terephthalate (PET), yield aromatic rings directly from their inherent structures. For instance, PS can recover up to 75% of its styrene at a relatively low temperature (450°C) and can attain 78.7% recovery at 600°C [58,59].

- Heteroatom-containing polymers, including those with oxygen, chlorine, fluorine, or nitrogen, follow diverse thermal decomposition pathways. These pathways often involve recombination reactions or the formation of stable molecules such as CO/CO_2, HCl, and NH_3. Nitrogen-containing polymers such as PU and PA undergo rearrangements at temperatures in the range of 250°–450°C, with a certain likelihood of producing linear fragments with functional groups such as -amino ($-CH_2-NH_2$) and -nitrile (-CN) groups [60,61]. Chlorine-containing polymers (e.g., PVC) experience significant aromatization post-dehydrochlorination, as chlorine's electronegativity weakens bond energy, thereby promoting the formation of reactive sites. These sites can react with radicals or create unsaturated species precursors that lead to aromatic and polyaromatic structures [62,63].

In summary, the thermal decomposition of polymers in heterogeneous plastic waste is governed by their structural characteristics. These differences underscore the need for a generalized representation that captures the chemical environments rather than individual polymer identities. The Carbon Bond Group (CBG) framework introduced next provides such representation by mapping chemical composition into bond-level categories, enabling correlations between structural features and product distributions across diverse mixtures.

4.2.2. Carbon Bond Group (CBG) Model

From a CN perspective, steam cracking of plastic waste can be seen as a process that involves two high-dimensional systems: the feedstock and the product slate. On the product side, the cracking generates a large number of chemical species, each with its own yield. If each species is treated as a variable with domain $\in \mathbb{R}^+$, the entire set constitutes a high-dimensional CN. Similarly, the feedstock space becomes high-dimensional when each polymer type in a heterogeneous mixture is treated as an independent variable. To address this complexity, the aim here is to introduce dimensionality reduction transformations explicitly linked to the system's structure, in order to compress the information while preserving or even enhancing chemical interpretability.

Following the approach introduced in **Paper II**, carried on in **Paper V**, and extended in **Paper VI**, instead of describing the feedstock in terms of its full polymeric composition, which may involve many different plastics, each with distinct structures, the composition is projected into a bond-based coordinate system denominated the Carbon Bond Group (CBG) classification framework. In its conceived form, polymers contained in the feedstock are categorized based on a three-group system that reflects their carbon structures: 1) carbons bound to heteroatoms (C-xE) by x number of bonds, where E represents heteroatoms such as O, Cl, N, etc.; 2); carbons in aliphatic bonds (C-AL), which includes paraffinic and olefinic structures; and 3) carbons in aromatic bonds (C-AR), accounting specifically for aromatic rings. Figure 9 shows the ratio of each carbon-bond basis with respect to the carbon of different polymers' repeating unit.

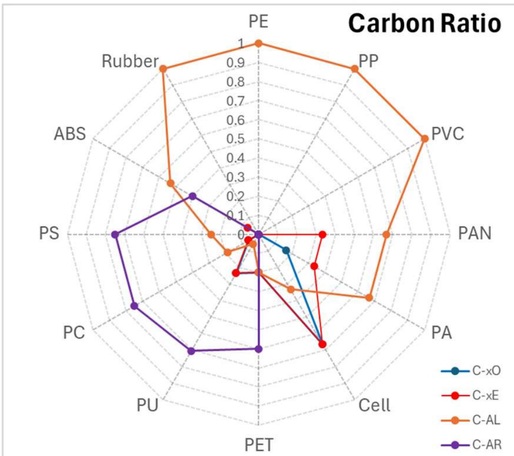


Figure 9. Transformation of polymers into the carbon bond groups framework

The general approach, introduced in **Paper II**, can take a simplified form for feedstocks with low heteroatom content other than oxygen. In such cases, the heteroatoms bond group (C-xE) is reduced to consider only carbons attached to oxygen by single or double bonds (C-xO), the remaining is assigned to C-AL. Based on observations of **Paper V**, a special case is set for ester-containing structures where both oxygens are assigned to the same carbon (e.g. in PET).

Based on this framework, **Paper VI** advances the qualitative carbon bond classification introduced in **Paper II** into a formal compositional representation. The bond groups are treated as structured state variables embedded in a constrained space, where each feedstock is represented as a normalized carbon-fraction vector subject to physical admissibility conditions. In this formulation, the CBG vector elements define not only a reduced description of feed structure, but also a mathematical scaffold in which conservation laws and other constraints can be explicitly enforced. This abstraction enables steam cracking to be expressed as a transformation between comparable compositional spaces, thereby providing the foundation for the predictive and condition-dependent mapping developed in **Paper VI**.

Formally,

- Let a feedstock's bond group composition be described by the vector $\boldsymbol{f} = [f_{C-xO}, f_{C-AL}, f_{C-AR}]^T$ where each component satisfies $f_b \in [0,1]$ and $\sum_b f_b = 1$ (super-index T means vector transposition). This vector, referred to as the feedstock composition vector lies within a feasible set Δ_2^F.
- The feasible set $\Delta_2^F = \{\boldsymbol{f} \in \mathbb{R}^3 \mid \mathbf{1}^T\boldsymbol{f} = 1,\ f_b \geq 0\}$ corresponds to a two dimensional probability simplex, a triangular slice of $\mathbb{R}^3$, also referred to as a ternary simplex space. Here, $\mathbf{1} = [1,1,1]^T$ corresponds to a column vector of ones. The set Δ_2^F is convex, meaning that any convex combination of feasible compositions remains feasible (Equation (20)). Points along the triangle's edges represent binary mixtures of bond groups, while the vertices correspond to pure bond group states. In this space, the triplet $(f_{C-xO}, f_{C-AL}, f_{C-AR})$ constitute barycentric coordinates within the simplex [1].
- The dimensionality reduction proceeds as follows: Let $\boldsymbol{x} \in \mathbb{R}_{\geq 0}^{Nc}$ denote the high-dimensional polymer composition of a feedstock composed of a set of different polymers $\{PE, PET, ..\}$, where N_c indicates the number of polymeric components. Then, the vector $\boldsymbol{x} = [x_{PE}, x_{PET}, \ldots]^T$ contains the normalized mass shares of each polymer and satisfies $\mathbf{1}^T\boldsymbol{x} = 1$. Let $\boldsymbol{D} \in \mathbb{R}^{3\times N_c}$ be a $3 \times N_c$ operator matrix, whose rows contain the carbon-bond group fractions for each polymer. The unscaled bond composition is then given by $\boldsymbol{Dx} \in \mathbb{R}_{\geq 0}^3$, and the normalized feedstock vector is obtained as:

$$\boldsymbol{f} = \frac{\boldsymbol{Dx}}{\mathbf{1}^T\boldsymbol{Dx}} = norm(\boldsymbol{Dx}) = [f_{C-xO}, f_{C-AL}, f_{C-AR}]^T \in \Delta_2 \tag{19}$$

Figure 10 shows an illustration of Equation (19) transformation:

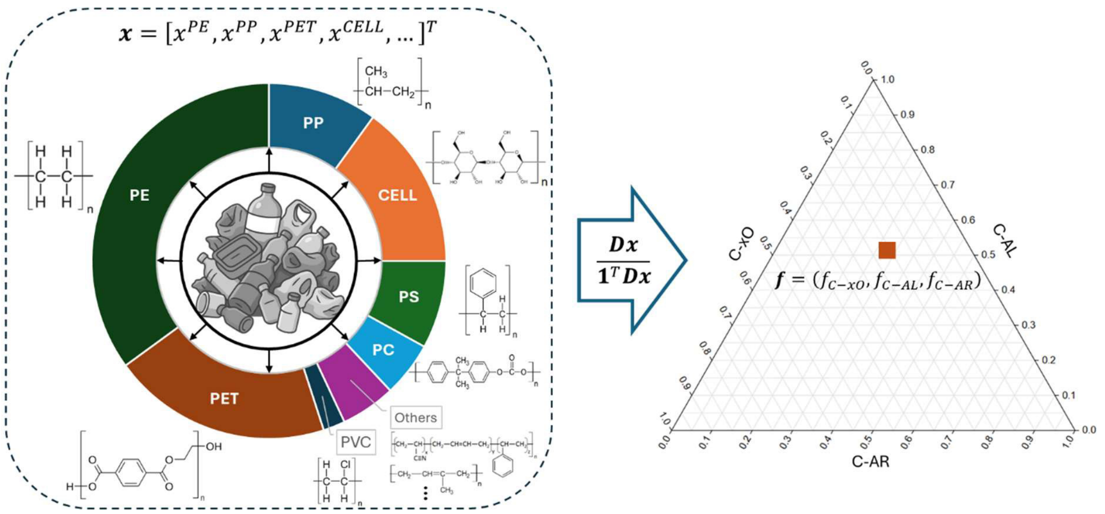


Figure 10. Schematic of CBG's dimensionality reduction transformation of a feedstock's polymeric composition.

[1] Barycentric coordinates are a way to represent a point within a triangle (a simplex) as a weighted average of the vertices. Consider a random point q in a triangle with vertices a, b, c. Then, q can be expressed as a barycentric combination of a, b, c by $q = ka + mb + nc$, where (k, m, n) are the barycentric coordinates constrained by $k + m + n = 1$.

Overall, this representation has several important properties. First, the reduced space enables straightforward visualization of feedstock diversity in terms of fundamental structural features, while preserving mechanistic interpretability (Figure 10). Second, mixtures of feedstocks can be expressed as convex combinations of their constituents, which ensures consistency with carbon mass balances. Formally, for two vector feedstocks defined as $\boldsymbol{f_1} = \left[f_{C-xO,1}, f_{C-AL,1}, f_{C-AR,1}\right]^T$ and $\boldsymbol{f_2} = \left[f_{C-xO,2}, f_{C-AL,2}, f_{C-AR,2}\right]^T$, their aggregation will produce a new vector $\boldsymbol{f_{mix}}$ obtained through a convex combination [1], as shown by Equation (20).

$$\boldsymbol{f_{mix}} = [f_1 \quad f_2] * W = \begin{bmatrix} f_{C-xO,1} & f_{C-xO,2} \\ f_{C-AL,1} & f_{C-AL,2} \\ f_{C-AR,1} & f_{C-AR,2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = w_1 f_1 + w_2 f_2, \quad \sum_i w_i = 1 \tag{20}$$

Where W represents the mixing operator that contains the normalized shares of the feedstock vectors $\boldsymbol{f_1}$ and $\boldsymbol{f_2}$ within $\boldsymbol{f_{mix}}$, See Figure 11.

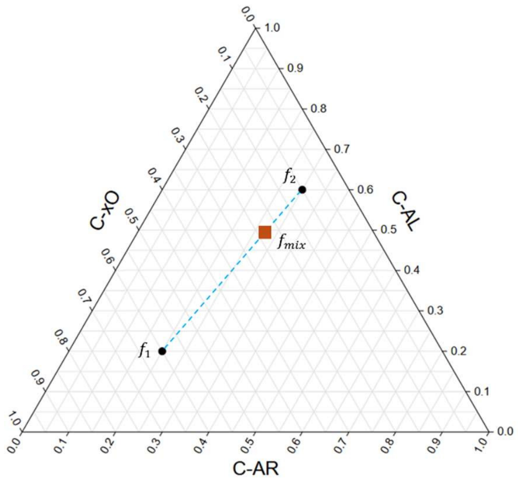


Figure 11. Example of two composition vectors aggregation in the ternary simplex Δ_2 of the feedstock space.

On the process output side, the product species set can also be projected onto three fundamental condensed basis groups: COx, aliphatics, and aromatics (see Figure 12), to create an analogous reduced output space Δ_2^P. Soot or Char obtained from the conversion is accounted within the aromatics group. Contributions of carbon to the product slate from external char/soot sources are considered negligible.

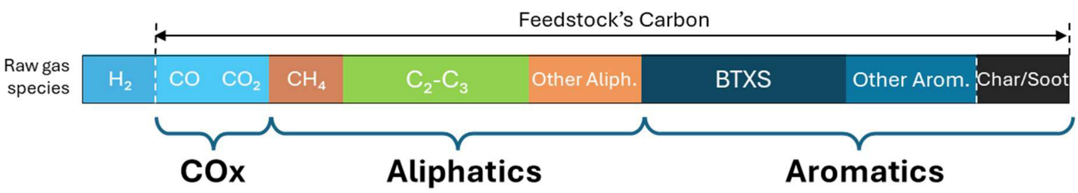


Figure 12. Product Groups Species Classification (See Section 5 for further details on the characterization methods used in this work).

[1] A convex combination y corresponds to the linear combinations (or weighted average) of points or vectors where all the coefficients (weights) are non-negative and sum to one, i.e., $y = \alpha_1 x_1 + \alpha_2 x_2$, $\alpha_1 + \alpha_2 = 1$

In this way, the cracking process can be conceptualized as a transformation between two lower-dimensional, chemically meaningful composition spaces (See Figure 13): from feedstock bond composition to product group distribution. This mapping provides a tractable basis for analyzing correlations between feedstock structure and product distributions.

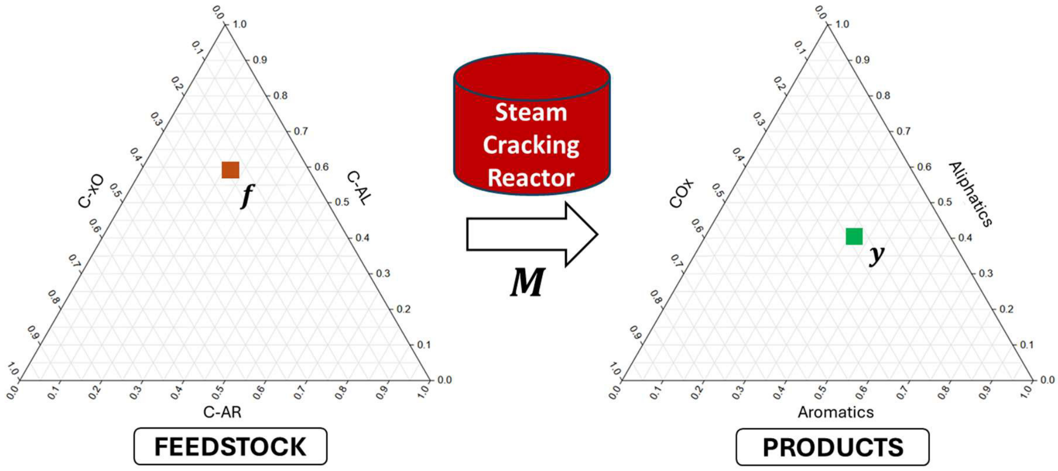


Figure 13. Representation of the transformation between component vectors feedstocks to products in their respective ternary simplex space ($F \rightarrow P$) under the steam cracking process.

Formally, the transformation can be expressed as a function $M(\boldsymbol{f})$, mapping elements from the feedstock simplex to the products simplex, i.e., a simplex-to-simplex-map as presented in Equation (21).

$$M: \underbrace{\Delta_2^F}_{feedstock} \rightarrow \underbrace{\Delta_2^P}_{products} \tag{21}$$

Representing heterogeneous mixtures in this reduced yet interpretable way the carbon bond-based framework can provide qualitative insights into the influence of feedstock structure over process outcomes. More importantly, it enables modelling the reactor as a mathematical transformation between two low-dimensional spaces, which can be used for prediction and for embedding structural constraints into data-driven models. These constraints can act as additional layers in a CN instance, strengthening predictive frameworks by incorporating chemical structure properties directly into their formulation. This approach extends the concept of physics-informed dimensionality reduction, complementing the PSM introduced in Section 4.1 and setting the stage for the hybrid data-model integration strategies discussed in Section 4.3.

4.3. Machine Learning Integration into CBG Model Framework

4.3.1. Mapping Feedstock and Products Spaces

As discussed in section 4.2, the dimensionality reduction achieved by the carbon bond groups framework, defines two simplex spaces: one for feedstock and one for products. The elements of these spaces will be defined as composition vectors in the simplex Δ_2, in the form described by Equation (22).

$$\boldsymbol{f} = [f^{C-xO}, f^{C-AL}, f^{C-AR}]^T$$
$$\boldsymbol{y} = [y^{COx}, y^{aliph}, y^{arom}]^T \tag{22}$$

Let's assume $M(\boldsymbol{f}) = \boldsymbol{y}$ as a fixed operative regime linear map from feed bond composition to product groups, such as:

$$\boldsymbol{y} \approx \boldsymbol{K}\boldsymbol{f}, \qquad with\, \boldsymbol{K} \in \mathbb{R}^{3x3}_{\geq \mathbf{0}}, \mathbf{1}^T\boldsymbol{K} = \mathbf{1}^T \tag{23}$$

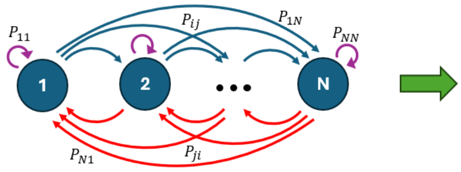
The interpretation of Equation (23) is that for each bond type b, the column $K_{:b}$ gives the distribution of that bond's carbon among $\{COx, aliph, arom\}$. The elements K_{gb} must be larger than zero ($K_{gb} \in \mathbb{R}_{\geq 0}$), and the system must enforce carbon partitioning, i.e., the sum of the elements of each column must be equal to one. With these conditions, $\boldsymbol{K}$ takes the form of what is known as a column-stochastic matrix defined by equation (24), where each column corresponds to a probability distribution over the product groups from each bond group type.

$$\boldsymbol{K} = \begin{bmatrix} K_{COx,C-xO} & K_{COx,C-AL} & K_{COx,C-AR} \\ K_{aliph,C-xO} & K_{aliph,C-AL} & K_{alip\ ,C-AR} \\ K_{arom,C-xO} & K_{arom,C-AL} & K_{arom,C-AR} \end{bmatrix} \tag{24}$$

The elements K_{gb} in equation (24), represent a probability of conversion (or conversion index) between bond b to product group g. Specifically, these probabilities can be interpreted as eventual outcomes from multistep, long-run absorption probabilities from the stochastic process occurring between transient states of the bond groups. This is a concept that will be developed further in the next section.

4.3.2. Transition Matrix and Absorbing Markov Chains

Formally, a Markov chain is a mathematical model that describes a system that moves between different "states" over time in a probabilistic way. The key idea is that the next state depends only on the current state (not on the history of previous states), and the probabilities of transitioning from one state to another are fixed. These transition probabilities are often organized into a transition matrix P. By convention each row of P represents the current state, each column represents the possible next states, and the entries are the probabilities to pass from the current state to the next one (i.e. each row must sum to 1) [64,65]. The concept is shown in Equation (25).



$$P = \begin{bmatrix} P_{11} & \cdots & P_{1N} \\ \vdots & \ddots & \vdots \\ P_{N1} & \cdots & P_{NN} \end{bmatrix}; \quad \sum_{j=1}^{N} P_{ij} = 1 \tag{25}$$

Now, let's consider an absorbing Markov chain. An absorbing Markov chain is a special type where some states are "absorbing". An absorbing state is the one where, once the system enters it, it stays there forever, i.e., the probability of transitioning to itself is 1, but to anywhere else is 0. In general, a Markov chain is absorbing if:

- There is at least one absorbing state.
- From every non-absorbing (transient) state, it's possible to eventually reach an absorbing state.

Conversely, transient states are those where the system can leave and enter back with a certain probability. However, over time, the chain will get "absorbed" into one of the absorbing states with probability 1.

In our concerned case, as discussed in Section 3.2, the thermochemical conversion involves a large series of stochastic free radical reaction events happening in the reactor, transient in nature, which end up over time into condensed molecules which form the observed species distribution. Therefore, keeping the parallelism for the CBG framework, the state sets will be:

- Transient (input) states $\{C - xO, C - AL, C - AR\}$: where a carbon "starts" labeled by its bond group (carbons bound to oxygen, aliphatic, aromatic).
- Absorbing (output) states $\{COx, aliph, arom\}$: once a carbon is counted in a product group, it stays there.

In order to analyze absorbing chains, the transient and absorbing states in matrix P are rearranged so transients come first, then absorbings. In general, for a fully absorbing Markov chain with t transient states and r absorbing states, the transition matrix P in its canonical form will be described as Equation (26) [65].

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix} \tag{26}$$

Where:

- Q: $t \times t$ matrix of transitions between transient states.
- R: $t \times r$ matrix of transitions from transients to absorbings.
- 0: $r \times t$ matrix of zeros (can't leave absorbings).
- I: $r \times r$ identity matrix (stuck in respective absorbing state).

In the CBG model, both t and r will be equal to 3. The matrix Q contains the probabilities of transitions among bond groups before finalization. It allows to represent in-process relabeling from rapid transient-to-transient steps (e.g. certain cyclization tendency that may transform C-AL into C-AR within the radical cloud) before the bond group is ultimately assigned to a product group. This representation can mimic a complex radical soup where very short-lived intermediates adjust the "label" prior to final accounting. On the other hand, R will contain the probabilities to go from the "transient" bond groups to the "absorbing" product groups in one step.

It is worth to note that in a general Markov chain, by definition, the bottom-left matrix of P accounts for transitions from absorbed states into transient ones, while the bottom-right will account for transitions between absorbed states. These blocks are not necessarily zero and identity matrices, respectively. However, in a fluidized bed, the reaction conditions are such that once the product species mixes with steam and leave the bed there are no further free radical reactions occurring. Overall, the reactor configuration and the presence of steam as a dilution agent restricts the runaway of secondary reactions from proceeding further [66]. Therefore, as a first-order approximation, it is reasonable to assume a zero matrix for the bottom-left block (indicating no transitions back to transient states after condensation) and an identity matrix for the bottom-right block (indicating that condensed species remain stable without interconversion). This assumption aligns with the standard absorbing Markov chain formulation in Equation (26), which, for this system, can be schematically represented in Figure 14.

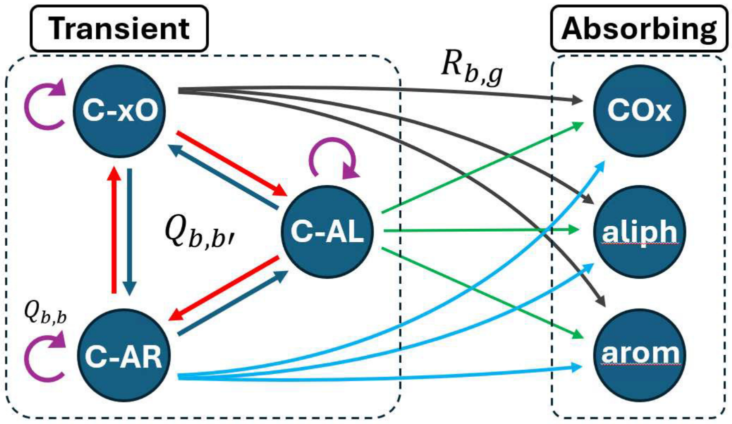


Figure 14. Schematic of the absorbing Markov chain in the CBG Model. $Q_{b,b'}$ indicates the probabilities between transient states and themselves, $R_{b,g}$ indicates probabilities from transients to absorbing states

It is important to clarify that the CBG model works based on a generalization framework of the key chemical environments present in the inputs and outputs of the reaction system. A more detailed radical-level model of the system might have to incorporate into the matrices all kinds of possible transient states from the large set of radical fragments, as well as the complete slate of condensed species obtained out from the reaction mechanisms from the steam cracking. Such a model can rapidly increase in complexity, and become challenging to solve, which is why dimensionality reduction strategies (such as the CBG model) must come into place at some point.

Now, given an absorbing chain, the probability to be absorbed into a given product in exactly one step is R. In two steps, it is QR (one transient step followed by absorption). In three steps, it becomes Q^2R, and so forth. Considering all possible steps that the system can take to reach its end, the probability of the eventual absorption will be thus described as an infinite geometric-like series. Such series is convergent and follows Equation (27) [65].

$$R + QR + Q^2R + \cdots = (I + Q + Q^2 + \cdots)R = \frac{1}{I - Q}R = NR = B \tag{27}$$

Matrix N is commonly known as the fundamental matrix. Its entries N_{ij} represent the expected number of times the chain will visit a transient state j from the transient i before reaching an absorbing state (bond to bond transition). It denotes all the pathways the system can wander among transient states prior to absorption. Matrix B is known as the absorption probability matrix and its entries B_{ij} give the probability that the Markov chain, starting in transient state i, will eventually be absorbed in absorbing state j. Translating this to the CBG model, B has entries $B_{b \to g}$ which gives the probability that a carbon starting in bond group b ends in product group g. This matches the definition for the elements of **K**. Since B corresponds to a row stochastic matrix, then **K** will be described by:

$$\boldsymbol{K} = B^T \tag{28}$$

Thus, $\boldsymbol{K}$ can be considered as the reactor's absorption map, a transformation from the feed simplex to the product simplex, representing the "end result" after all stochastic radical steps have condensed out. Each column of $\boldsymbol{K}$ is then a probability distribution over product groups for a given bond group and can be interpreted as the final product slate from the reactor if the feed were composed entirely of that bond group.

4.3.3. Machine Learning Integration

From Equations (27) and (28), matrix $\boldsymbol{K}$ can be defined at a fundamental level by the matrices Q and R. However, inference of Q and R is not unique since many pairs can give the same absorption $B = NR$. In practice, it may require the implementation of stochastic-type simulations, such as continuous-time Markov chain models [67], over the reaction network along with reaction rates from kinetic models, to get an estimation on their entries but this beyond the scope of the present study.

Alternatively, since the aim is to find $\boldsymbol{K}$, a data driven approach can be implemented to get meaningful estimations of its entries directly from experimental results. Based on the concepts presented in Section 2.1, the entries of matrix $\boldsymbol{K}$ can indeed be seen as the network's variables of a CN with a domain between 0 and 1. Furthermore, $\boldsymbol{K}$ corresponds to the linear mapping between simplex spaces from the feedstock composition vector $\boldsymbol{f}$ to the products $\boldsymbol{y}$. Therefore, given an experimental dataset $\{\boldsymbol{f}^{(n)}, \boldsymbol{y}^{(n)}\}$, the transformation is expressed as:

$$y^{(n)} = Kf^{(n)} \tag{29}$$

For N datasets, Equation (29) becomes a linear equation system where $\boldsymbol{K}$ can be found as the joint constrained fit from the optimization of the regression problem with a loss function of the form:

$$\min_K \sum_n^N \left\| y^{(n)} - Kf^{(n)} \right\|_2^2 + \lambda \|K\|_2^2 \tag{30}$$

Here $\|\cdot\|_2$ corresponds to the Euclidean norm (L2 norm for vectors and Frobenius norm for matrices). The parameter λ comes as a tunable regularization term typical of Ridge regression[1], which is especially useful to enhance model's generalization for small datasets [68]. This COP instance is part of a classical machine learning algorithm to learn linear mapping functions between inputs and outputs.

In addition to data fitting, domain knowledge must be incorporated through additional constraints from the physical and probabilistic priors that define the conversion process, such as:

[1] Ridge regression is a statistical technique that enhances linear regression by adding a penalty term to the loss function. Also known as L2 regularization, it helps to prevent overfitting by "shrinking" the coefficients towards zero, reducing model complexity, and improving generalization by mitigating issues like multicollinearity (highly correlated predictor variables). λ controls the penalty's strength, higher values lead to more shrinkage.

- Nonnegativity: $K_{gb} \geq 0$
- Carbon partitioning (column-stochastic condition): $\sum_g K_{gb} = 1$
- Physics tendency inequalities: $K_{arom,C-AR} \geq K_{arom,C-AL}$, $K_{COx,C-xO} \geq K_{COx,C-AL}$, $K_{aliph,C-AR} \leq \epsilon$

(31)

With the last constraint, the aim is to make the diagonal terms free to be large if the data supports it, while the off-diagonal terms will capture plausible cross-bond to product transfers (e.g. negligible $C - AR \rightarrow aliph$ transfer).

Together, these constraints form a full PICN model, and are intended to reduce the solution space to physically meaningful boundaries to enhance the numerical stability of the learning process and improve the results interpretability especially when dealing with small datasets (see Section 2.3).

This PICN can be solved using convex optimizer algorithms (such as CVXPY in python as applied in **paper II**) or via metaheuristic algorithms such as PSO, with convex solvers generally being faster.

Finally, the predicted product group y_g for $g \in \{COx, aliph, arom\}$ will be a linear predictor of the form:

$$y_g^{pred} = K_{g,C-xO} f_{C-xO} + K_{g,C-AL} f_{C-AL} + K_{g,C-AR} f_{C-AR} \tag{32}$$

Model validation can be performed by computing the R^2 coefficient and comparing predicted versus measured values using cross-plots. If enough data is available, k-fold techniques can be applied to evaluate generalization (out-of-sample performance) of the model.

4.3.4. Generalization of CBG model for Variable Operative Conditions

In practice, the reactor's operative conditions exert a decisive influence on the observed transformations. Temperature (T), steam dilution (S/C), residence time (τ), Bed Material Circulation (BM) among others, have an impact on the balance between primary scission and secondary reactions, thereby altering the effective probabilities of conversion across bond groups. To capture this dependency, the matrix **K** can be formulated as a function of the operative conditions, $\boldsymbol{K}(\boldsymbol{u})$, where $\boldsymbol{u} \in \mathbb{R}^m$ denotes a vector of relevant operative conditions. The mapping in Equation (23) can be now generalized as:

$$\boldsymbol{y}(\boldsymbol{u}) = \boldsymbol{K}(\boldsymbol{u})\boldsymbol{f} \tag{33}$$

To preserve simultaneously the fundamental structural constraints of nonnegativity and column stochasticity, each column of $\boldsymbol{K}(\boldsymbol{u})$ can be parameterized using a SoftMax transformation applied to a set of logits $Z_{gb}(\boldsymbol{u})$:

$$Z_{gb}(\boldsymbol{u}) = A_{gb} + B_{gb}^T \phi(\boldsymbol{u}); \qquad K_{gb}(\boldsymbol{u}) = \frac{\exp\left(Z_{gb}(\boldsymbol{u})\right)}{\sum_{g'} \exp\left(Z_{g'b}(\boldsymbol{u})\right)} \tag{34}$$

Here, $Z_{gb}(\boldsymbol{u})$ represents an unconstrained latent score (logit), and it can be interpreted as a measure of the relative tendency of a carbon atom in group b to transition into group g under conditions $\boldsymbol{u}$.

The softmax function then transforms via nonlinear map these scores into probabilities, ensuring all entries are strictly positive and each column sums to one. This transformation guarantees that $\boldsymbol{K}(\boldsymbol{u})$ remains column-stochastic and thus structurally admissible at all times throughout the learning process.

The term $\phi(u)$ represents a feature map assembly of operative conditions. Depending on the level of detail desired, it may take the form of: simple raw variables or polynomial expansions (e.g. $\phi(u) = [T]$ or $\phi(u) = [T, T^2]$), logarithmic or interaction terms (e.g. $\phi(u) = [T * log(S/C), T * \tau, \dots]$), or more flexible basis function such as a spline.

In this work, this formulation is presented as a proof of concept in Section 6.2.2, where only the reactor's temperature is considered as the dominant operative variable. However, the structure of Equation (34) can accommodate richer feature maps for $\phi(u)$ if additional conditions are incorporated.

This construction is closely related to what is known as the Generalized Linear Model (GLM) widely used in statistics and machine learning [69]. In a GLM, the expected value of a response variable is linked to a linear combination of input features via a nonlinear link function. Here, SoftMax acts as such link, mapping the linear predictors $Z_{gb}(\boldsymbol{u})$ into a valid probability distribution across product groups (See Figure 15). This is equivalent to a multinomial logistic regression, ensuring both interpretability and compatibility with standard optimization techniques.

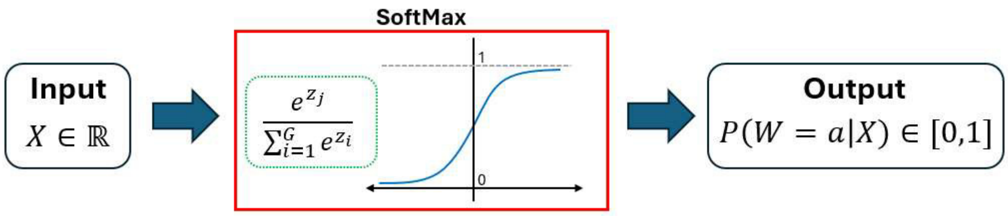


Figure 15. SoftMax transformation of variable X into a probability distribution (adapted from [70])

Now, the learning problem for Equation (33) will be defined by a loss function that must balance fidelity to observed data with regularization of the parameters and the penalties of the physics priors (soft constraints):

$$L = L_{data} + L_{reg} + L_{phys} \tag{35}$$

$$\Rightarrow L = \sum_{n} \left\| y^{(n)} - K\left(u^{(n)}\right) f^{(n)} \right\|_2^2 + \left(\lambda_A \, \|A\|_2^2 + \lambda_B \, \|B\|_2^2\right) + \sum_{h} w_h \left(\Delta^{(h)}_{K_{i,j}(u^{(n)})}\right)^2 \tag{36}$$

The first term enforces data fit through the Euclidean norm (L_{data}). The ridge regularization terms (L_{reg}) penalize numerically over-complex parameterizations. The last term (L_{phys}), sets violation penalties weighted by the vector w_h over the set of domain-related priors such as the chemical tendency inequalities of Equation (31), e.g. $\Delta^{(h)}_{K_{arom,\cdot}} = \max\{0, \left(K_{arom,C-AR} - K_{arom,C-AL}\right)\}$.

Since there is a nonlinear transformation involved due to use of the SoftMax link, the optimization can be performed through PSO. Nonetheless, robust and generally faster gradient descent algorithms as ADAM can also be implemented considering the data for N samples and the respective gradients of the parameters involved [71].

Formally, given the squared error, the gradient with respect to $K_{gb}\left(u^{(n)}\right)$ at sample n is:

$$\frac{\partial L}{\partial K_{gb}(u^{(n)})} = 2\left(\hat{y}_g^{(n)} - y_g^{(n)}\right)f_b^{(n)} + \sum_h 2w_h\Delta_K^{(h)}\frac{\partial\Delta_K^{(h)}}{\partial K_{gb}}$$

For a fixed column b, the SoftMax Jacobian will be:

$$\frac{\partial K_{gb}}{\partial Z_{hb}} = K_{gb}\left(\delta_{gh} - K_{hb}\right)$$

With δ_{gh} as the matricial Kronecker delta (one if $h = g$, zero otherwise). So,

$$\Rightarrow \frac{\partial L}{\partial Z_{hb}} = \sum_g \frac{\partial L}{\partial K_{gb}(u^{(n)})} K_{gb}\left(u^{(n)}\right)\left(\delta_{gh} - K_{hb}\left(u^{(n)}\right)\right)$$

Finally, since $Z_{hb}(\boldsymbol{u}) = A_{hb} + B_{hb}^T\phi(\boldsymbol{u})$, the gradients for A and B for the algorithm implementation will be defined by:

$$\frac{\partial L}{\partial A_{hb}} = \sum_n \frac{\partial L}{\partial Z_{hb}(u^{(n)})}, \quad \frac{\partial L}{\partial B_{hb}} = \sum_n \frac{\partial L}{\partial Z_{hb}(u^{(n)})}\phi\left(u^{(n)}\right) \tag{37}$$

Alternatively, specialized machine learning libraries as Pytorch or TensorFlow in Python can offer a smooth and straightforward implementation of ADAM and other gradient descent algorithms via their in-built methods and optimization functions [72,73].

Overall, this model's extension enables the CBG model to move beyond static correlations towards a condition-responsive mapping, where not only feedstock structure but also operative reactor conditions jointly determine predicted outcomes. This broadens the model's predictive scope and opens the possibility of incorporating process optimization into the same constraint-based formulation.

Additional expansions to the CBG model formulation can be explored such as: a generalized reactor-type dependent model, radical-based transient groups decoupling to evaluate specific radical such as methyl, backward polymeric composition estimation and integration of CBG model and PSM into a general physics informed neural networks (PINNs). These are beyond the scope of the present document but will be overviewed in Section 7.

5. Experimental Setup

5.1. Sampling Setup and Carbon Balance

The experiments evaluated in this study were conducted at the Chalmers Power Central facility, where a DFB Gasifier/Cracker is coupled with a Circulating Fluidized Bed combustor that runs with biomass wood chips. Silica sand was used as the bed material and the reactor was fluidized with steam. The feedstock flows for the cracker processes are in the range of 40–160 kg/h of plastic materials, while the CFB is operated with flows in the range of 1,500–3,000kg/h of wood chips. A small flow of high-purity helium is added along with the steam (~150 kg/h), to serve as a tracer gas for the species quantification. The DFB configuration allows the continuous removal of the produced char, along with the bed material, transporting it back to the combustor side. Following conversion of the feedstock, a sample stream is continuously extracted at the reactor's exit and the remaining product gas is conducted back to the combustor (see Figure 16).

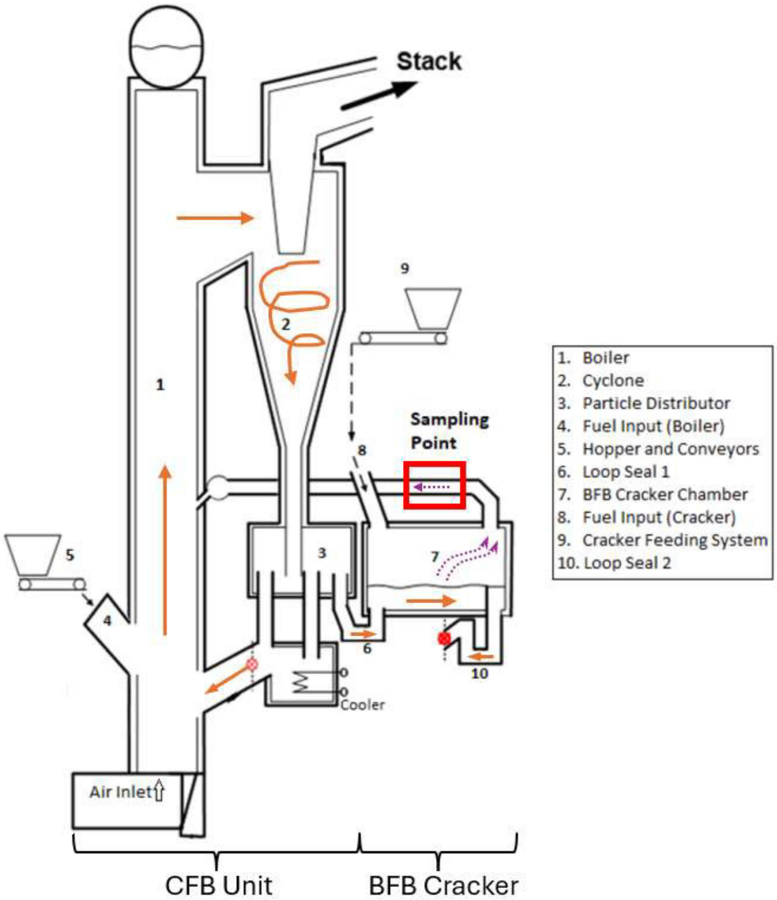


Figure 16. Schematic of the Chalmers power plant and gasifier. Orange Lines: Bed Material Circulation, Purple Lines: Producer Gas (modified from [74]).

The sampled stream passes through a high-temperature (~350°C) particle filter and then splits into two parallel streams (see Figure 17). The first stream passes through an isopropanol quenching loop, to remove condensable species such as water, long hydrocarbon chains, and aromatics. Further cooling is performed in a chiller, to ensure that no species condense downstream. After conditioning, the gases are pumped into a Micro-Gas Chromatograph Varian CP4900 (GC1) equipped with a Thermal Conductivity Detector (TCD). The micro-GC has two channels and uses

Molsieve 5Å (MS5Å) and PoraPLOT Q columns with argon and helium as carrier gases, respectively. Permanent gases and C1–C3 hydrocarbon species can be characterized online by sampling from the continuous gas stream at 3-minute intervals. A weekly calibration of the chromatograph is performed using five different concentrations of the expected species in the dried gas (H_2, He, N_2, O_2, CO, CO_2, CH_4, C_2H_2, C_2H_4, C_2H_6, C_3H_6 and C_3H_8).

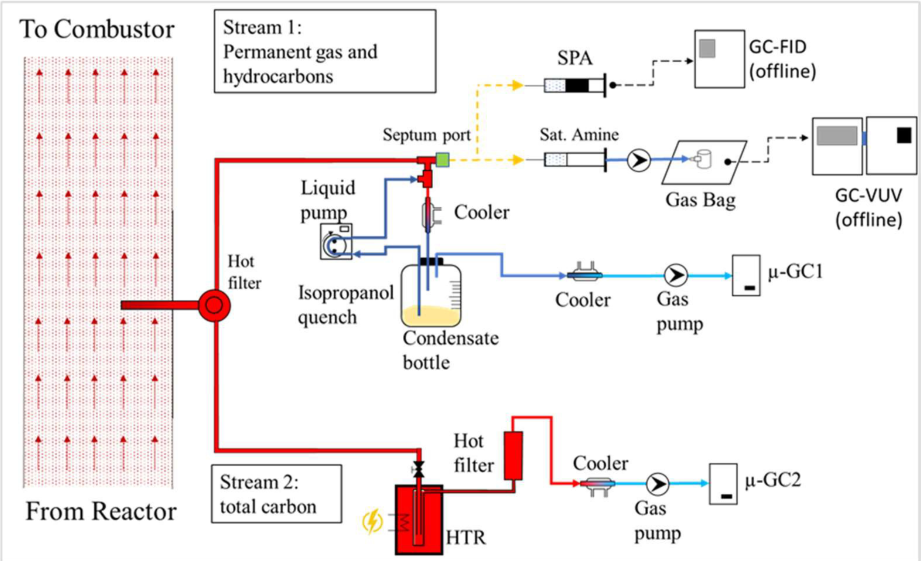


Figure 17. Sampling setup to characterize the permanent gases, hydrocarbons (aliphatics and aromatics) and total carbon in the reactor gas flow.

C4+ species are characterized in a GC-VUV. From a septum port located before the isopropanol loop, a gas sample is drawn through a saturated amine that acts as a filter for water and is collected into a gas bag at room temperature for characterization in the GC-VUV. The carrier gas is hydrogen, and nitrogen is used in the detector to make the gas cell inert. The chromatograph column used for the experiments is detailed in **Paper III**.

Regarding this last setup, the study presented in **Paper III** focuses on establishing a formal methodology for sampling and characterization based on GC-VUV. This methodology provides a comprehensive carbon balance for the cracking products slate within a reasonable timeframe, with the goal of future implementation for online characterization. The work showcases different methodologies, comparing the proposed sampling method with more-traditional methods in terms of analytic timeframe and level of species resolution. The evaluated methodology based on GC-VUV achieved optimal quantification for C4–C18 species in considerably less time than the more-conventional techniques, such as those based on solid phase adsorption described next.

From the same sampling port, a Solid Phase Adsorption (SPA) method is applied to obtain the aromatic fraction. In this method, the gas is subjected to suction with a 100 ml syringe at a constant rate, forcing it to pass through an adsorbent column (Supelclean ENVI-Carb/NH_2 SPE columns), which consists of an amine adsorbent layer (500 mg) followed by an activated carbon layer (500 mg). The adsorbed aromatics in the SPA column are subsequently eluted into a vial that contains a mixture of dichloromethane, isopropanol, and acetonitrile (8:1:1) using hexylbenzene and 4-ethoxyphenol as internal standards at concentrations suitable for the species quantification (~12,000 mg/L and ~250 mg/L, respectively). The vial is analyzed in a Bruker GC430 GC coupled with a Flame Ionization Detector (FID) and equipped with a mid-polar BR-17 MS (BR85877) column using H_2 as the carrier gas. Twenty-eight different aromatic species are quantified with boiling points ranging from monoaromatics, such as BTXS, to polyaromatics, such as naphthalene, anthracene, and triphenylene. More details regarding this method and the measured species can be found elsewhere [75].

The second hot stream flows through a HTR that is electrically heated to ~1,700°C for the complete reforming of all hydrocarbon species by the steam contained in the raw gas. After leaving the reactor, the gases are filtered to remove soot and cooled down to remove excess water before being pumped into another micro-GC Varian CP4900 (GC2). The gases are expected to be almost 100% pure syngas (H_2 , CO and CO_2) containing the corresponding amount of helium, which allows one to estimate the total carbon in the producer gas and to derive an indirect estimation of the char yield. The efficacy of the reforming process is determined according to the amount of methane detected (close to zero). The HTR operates continuously and in parallel with the remainder of the described sampling process. After completing the measurement set, the HTR reactor is flushed with air to burn out any particles of soot that may have formed during the reforming process, and the corresponding carbon amount is measured in terms of the CO_2 produced. This setup and method, developed by Israelsson et. Al. [55,74], provides a robust approach to estimate and validate the carbon balance of the sampled gases from the reactor (Figure 18).

The solution of the carbon balance from the reaction system based on the characterization setup is depicted in Figure 18.

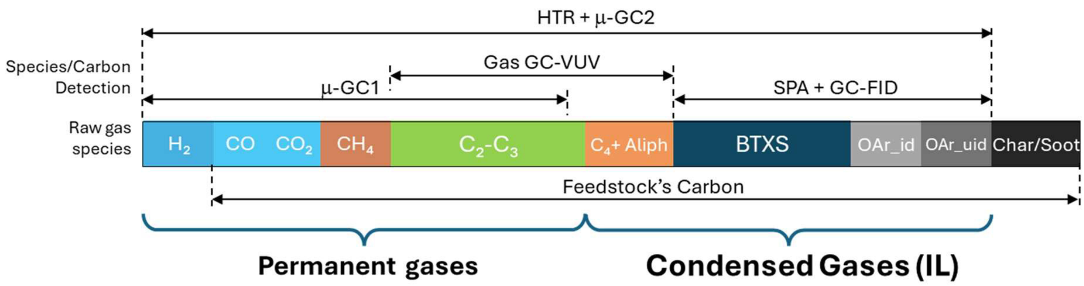


Figure 18. Carbon Balance yields estimation from the implemented characterization setup. AR_kn refers to other known aromatics (e.g. polyaromatics) whereas Ar_ukn refers to other unknown aromatics also identified by GC-FID from the SPA elution.

5.2. Feedstocks and Operative Conditions

Table 2 shows the materials used in the development of the results of the present work, with a brief description of their origin, overall relevant characteristics and polymeric composition set.

Material	Pol. Types	Chemical Characteristics	Description
Polyethylene (PE)	PE	- High polyolefin content - Very low oxygen/aromatic content - None/very low ash content	Virgin Polymer Pellets.
Mechanical Recycled Polyolefins (MRP)	PE,PP	- High polyolefin content - Very low oxygen/aromatic content - None/very low ash content	Virgin Polymer Pellets.
Plastic Packaging Sorting Reject. 2D Packaging (PSR2D)	PE, PP, PET, Cell, PA, PVC	- Medium-high polyolefin content - High oxygen content - Medium-low ash content	Reject from post-consumer sorted plastic packaging waste. Mixture of 2D packaging, e.g., multilayer films.
Plastic Packaging Sorting Reject. 3D Packaging (PSR3D)	PE, PP, PET, PS, PA, PVC	- Medium polyolefin content - High aromatic - Medium-low ash content	Reject from post-consumer sorted plastic packaging waste. Mixture of 3D packaging, e.g., bottles, food, cleaning and household recipients.
Mixed Packaging Waste (MPW)	PE, PP, PET, Cell, PVC (others: PU, PA, PS)	- Medium polyolefin content - High oxygen content - Medium ash content	Post-consumer unsorted packaging plastic waste.
CardBoard Recycling Residue (CRR)	PE, PP, Cell, PET, PVC (others: PU, PA, PS)	- High aliphatic carbon content - High oxygen content - Medium ash content	Post-consumer shredded stream of multilayer cardboard/plastic for food packaging.
PolyesterTextiles (TXTp)	PET, PA, PAN (Others: Wool, PVC, PU)	- Complex polymer blends - Low aliphatic carbon content - High oxygen content	Household textile waste after sorting the useful pieces of cloth.
Cotton-based Textiles (TXTc)	Cell, PET, PA, PAN (Others: Wool, PVC, PU)	- Complex polymer blends - Low aliphatic carbon content - High oxygen content	Textile waste after sorting the useful pieces of cloth.

Table 2. Heterogeneous materials used in this work with their respective description and general characteristics. PA, polyamide; PAN, polyacrylonitrile; PC, polycarbonate; PE, polyethylene; PP, polypropylene; PS, polystyrene; PU, polyurethane; PVC, polyvinyl chloride; PET, polyethylene terephthalate.

With the exception of the pure polyolefinic materials (PE and MRP), the rest of the feedstocks correspond to heterogenous streams coming from rejected fractions of post-consumer recycled products. All heterogenous batches were shredded and converted into pellets for reactor feeding.

The ultimate analyses of the three batches are presented in Table 3.

Item	PE	MRP	PSR2D	PSR3D	MPW	CRR	TXTp	TXTc
C	85.71	85.71	74.90	75.43	65.03	60.60	60.93	51.95
H	14.29	14.29	11.73	10.76	10.33	9.00	5.54	5.95
O*	0.00	0.00	8.13	8.64	14.64	21.00	28.96	37.96
N	0.00	0.00	0.67	0.54	0.52	0.35	3.15	1.53
S	0.00	0.00	0.04	0.03	0.09	0.07	0.07	0.11
Cl	0.00	0.00	0.20	0.31	0.70	0.20	0.12	0.37
Ash	0.00	0.00	4.34	4.28	8.69	8.75	1.22	2.11
LHV	42.80	44.61	36.03	37.02	30.98	30.14	28.13	17.87

Table 3. Elemental compositions ($\%w_{dry}$) and respective percent errors (%Err) for the materials used. *Calculated by difference-; LHV, low-heating value (MJ/kgdry).

Estimates of the polymeric compositions of the batches were obtained through manual sorting and NIR spectrometry before pelletization. Additionally, to reduce the uncertainty, the numerical estimation method developed in **Paper II** was implemented which uses a convex optimization model built on elemental balances and low heating values. Figure 19 shows the respective estimated polymeric composition for the evaluated feedstocks.

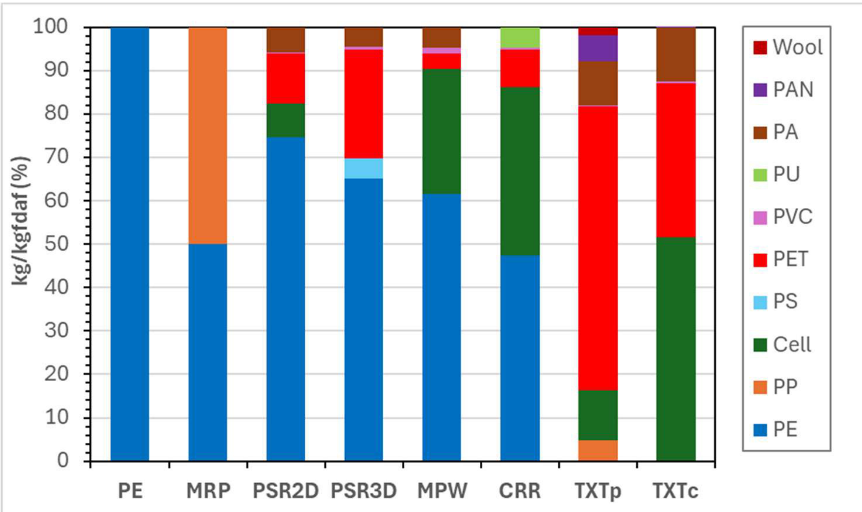


Figure 19. Feedstock Polymeric Composition (kgfdaf: kg of feedstock on a dried, ash-free basis)

The main operational conditions used in the DFB cracking reactor are indicated in Table 4. The cracker was fluidized with steam in a bubbling regime, the fuel residence time is estimated to be around 4–5 min, and the gas residence time 5–10 s. The column titled “Feeding position” refers to Figure 16, whereby the cracker can be fed through two different ports. Position 6 is situated at the top of loop seal 1, which consists of an extruder in which the feedstock, in pellet or granulate form, is compressed and heated to obtain a molten flow that pours down onto the cracker bed. In position 8, the feedstock, also in pellet or granulate form, falls by gravity into the bed via a set of rotary valves working in an airlock system [76].

Fuel	***Temperature Cracker (°C)***	***Bed material***	***Material Flow (kg_{daf}/h)***	***Feeding Mode***	***Feeding Position***	***Used in Section***
PE	780- 840	Silica Sand	~90	Molten flow via extrusion	6	6.1
PE	725-831	Silica Sand	63-100	Molten flow via extrusion	6	6.1/6.2
MRP	720-807	Silica Sand	60-75	Molten flow via extrusion	6	6.2
PSR2D	767-829	Silica Sand	49-58	Molten flow via extrusion	6	
PSR3D	760-826	Silica Sand	50-60	Molten flow via extrusion	6	
MPW	721-826	Silica Sand	50	Molten flow via extrusion	6	
CRR	742-804	Silica Sand	37-50	Molten flow via extrusion	6	
TXTp	735-815	Silica Sand	37-150	Top Feeding/ Extrusion	6,8	
TXTc	~756	Silica Sand	37-50	Molten flow via extrusion	6	

Table 4. Operative conditions in the BFB cracking reactor tested in this work

6. Results and Discussion

6.1. Parametric Model System for linear systems

This section presents the evaluation of the PSM applied to steam cracking of polyethylene under varying operational conditions. It builds on the concepts introduced on Section 4.1, where the PSM was formulated in detail. The discussion is based on findings from **Paper I** and begins by assessing the suitability of different probability distribution functions to represent carbon-number distributions of paraffins and olefins, including comparative analysis across severity levels and their implications for predictive robustness. Next, the section evaluates the model's capability for data quality assessment, demonstrating how the PSM can detect inconsistencies in experimental measurements by leveraging elemental balances and a parallel high-temperature reactor setup for total carbon estimation. Finally, the analysis expands the published work by evaluating the PSM against a broader dataset, incorporating experiments from Papers III and IV. This extension allows assessment of the model's robustness and interpretability, as well as exploration of how fitted parameters evolve with temperature, providing insights into the relationship between the process operative conditions and the parametric representation of product distributions. Overall, the results demonstrate the PSM's dual role as a predictive tool and a diagnostic framework, providing a structured basis for improving experimental reliability and supporting future integration into process control strategies.

6.1.1. Parametric System Model Evaluation

The functions in Table 1 were tested with the experimental results from the different operational conditions tested for the PE case (see Table 4). C1 to C4 groups were used as the observables for the PICN implementation. Figure 20 displays some of the experimental cases and the functions used from Table 1 to fit the results. In general, the topology of mono-parametric functions was found not sufficiently flexible to describe adequately the decaying behavior of the species. Hence, at least one additional shape parameter is needed to satisfy the conditions imposed. Therefore, the analysis is focused on the bi-parametric functions. Out of the five experimental cases studied in **Paper I**, only three cases are shown, for the sake of clarity, based on their relevance to the process severity (defined in this case as ethylene/propylene = 1/PER), and displayed from left to right as the low, medium and high severity cases respectively. The stacked bars in the upper row plots of Figure 20 show the experimental results for paraffin (orange) and the olefins (purple), as well as the different functions that were tested to describe the systems. The designation 'pfof' refers to the aggregation of the paraffin and olefin functions for each *k* group.

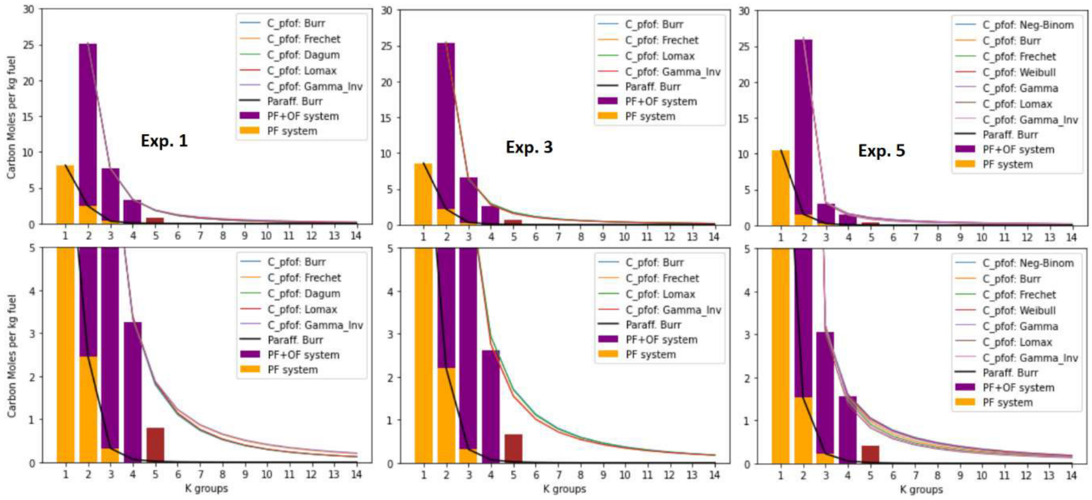


Figure 20. Parametric functions tested over measured paraffin and olefin species carbon distributions. Results in $molC/kg_{fuel}$ vs. k-group, obtained for different severity levels (low, medium, high from left to right). Bottom row: Zoomed-in plots. (Red column: incomplete group species measured).

Notably from Figure 20, the heavy-tailed functions, which include Burr, Fréchet, Lomax, and Inverse Gamma, proved to be the most-flexible in terms of fitting the data across all severity levels. These functions exhibit elongated tails over higher carbon groups, leading to a lower rate of decay and an enlargement effect on their total enclosed areas, as shown in Table 5. This characteristic is especially crucial in low-severity cases, where the yields of long-chain species are still relevant. The function's tail corresponds to the function's estimations over the C5+ groups, while its area their total estimated concentrations.

Table 5. Results for the total area, tail area (sum of C5+), and the parameters associated with the functions tested.

	Exp. 1				Exp. 3				Exp. 5			
Function	**Area**	**Area Tail**	**α1**	**α2**	**Area**	**Area Tail**	**α1**	**α2**	**Area**	**Area Tail**	**α1**	**α2**
Neg-Binomial	nan	nan	nan	nan	nan	nan	nan	nan	35.079	11.480	0.943	0.128
Con-Max	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
*Burr**	40.048	10.742	1.204	1.117	40.323	13.025	1.243	0.867	36.204	12.574	1.620	0.465
*Fréchet**	42.774	13.965	0.854	0.589	40.254	12.788	0.811	0.481	36.281	12.080	0.498	0.158
*Dagum**	40.077	10.789	0.442	1.399	nan	nan	nan	nan	nan	nan	nan	nan
Gompertz	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
*Weibull**	nan	nan	nan	nan	nan	nan	nan	nan	36.164	13.042	0.686	0.315
Gamma	nan	nan	nan	nan	nan	nan	nan	nan	34.500	10.594	0.066	0.126
*Lomax**	40.134	10.865	2.538	1.380	40.335	12.994	1.949	0.691	35.713	11.120	1.444	0.076
*Gamma_Inv**	43.458	14.631	0.760	0.429	40.438	12.940	0.744	0.342	36.122	11.622	0.439	0.061

*, Heavy-tailed functions; nan, No results.

At higher severities, the long chains progressively break down into shorter hydrocarbon species, contributing to the observed increase in data skewness. In such a scenario, topologies that feature rapid decay with relatively small estimated areas for the unmeasured groups, such as the Negative Binomial and Gamma functions (see Table 5), manage to accommodate the skewness of the data. The presence of such functions in the solution set also serves as an indication of a system getting closer to the ideal random breaking down scenario presented in Section 3.3. On the other hand, the heavy-tailed functions mentioned earlier remained adaptable across all cases, effectively capturing experimental data with high-level skewness while keeping the constraints, demonstrating their topology robustness to work in high- and low-severity scenarios.

Now, considering the hydrogen distribution, Figure 21 displays the same experimental cases as presented before. The bottom-row plots provide detailed views of the tested functions. The gray, blue, and purple-dashed curves represent the three hydrogen distribution scenarios: mono-enes (high), fully unsaturated (low), and fully conjugated (middle) cases of the olefin's gamma ratio effective function ($\gamma_{k,eff}^{of}$), respectively, as introduced in Equations (13) to (15). Notably, after the C7 group, most of the functions fell short of the allowed range, only Burr, Fréchet, and Inverse Gamma functions remained within the permissible region. These three functions consistently keep the hard hydrogen constraints in the whole domain and across experimental cases, which indicate their shape compatibility with the chemical characteristics of the olefin system; further enhancing the confidence in their respective predictions.

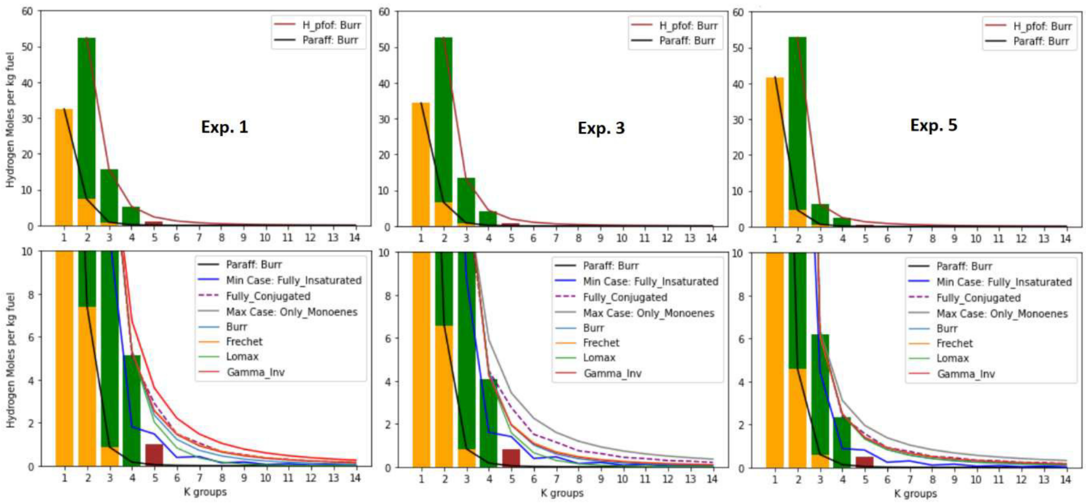


Figure 21. Results for the parametric functions applied to the species hydrogen distribution, expressed in $molH/kg_{fuel}$ vs. k-group. Upper row, Burr function results. Bottom row, zoom in and evaluation of the functions according with the gamma ratio effective. (Red column: incomplete group species measured).

From the results Burr was the sole function with a topology that was able to fit the data and all the system constraints across all the tested cases. Consequently, due to its robust performance, Burr is the function recommended for PSM applications and further testing under the evaluated process conditions.

In Figure 20 and Figure 21, the C5 species bar is indicated in red, denoting incomplete species data due to equipment limitations. Only two out of approximately fifteen potential olefin species could be measured with the characterization methodology used. In Figure 20, the mismatch is evident between the measures and the functions' estimations for the C5 group. In fact, regardless of the function considered, there is a noticeable gap of around 100% of the C5 bar's height. This effect is particularly pronounced in the low-severity cases. So, there is no solution that can fit the C5 measurements and at the same time fulfill the expected physical constraints of the system. Based on this, the study in **Paper I** underscores the model's predictive capabilities for the unmeasured carbon groups. At this point, the function's predictions remain physically consistent owing to the highly constrained framework built upon the set of chemical and mathematical considerations applied to the system. Thus, in the absence of adequate measurements for C5+ species, the PSM becomes an important and useful tool to get physically congruent estimations of the unmeasured carbon group species.

6.1.2. Parametric System Model Validation and Data Quality Assessment

As described in Section 4.1.3, the model outcomes can be validated by calculating the BOT. Such validation can be used in two ways. First, by checking the BOT calculated from the PSM estimations using Equation 16 align with the expected values given by Equation 18 using the results from the HTR reactor. Second, by deliberately introducing errors into the experimental results through modifications in the calibration curves. This dual validation strategy aims to assess both the accuracy of the model's estimations and its capabilities to evaluate the overall quality of the data by detecting errors in the measurements. Table 6 presents the results of the BOT calculation from the HTR outcomes versus the PSM estimations for both the calibrated and the mis-calibrated cases.

Table 6. Bed's oxygen transport (BOT) calculated from the HTR versus the value estimated from the model (PSM), for the calibrated case (left table) and the mis-calibrated case (right table).

Calibrated Case		
BOT HTR	BOT PSM	%diff
8.27	8.91	7.74
10.8	9.52	11.85
10.21	9.51	6.86
8.92	8.87	0.56
8.14	7.18	11.79

Mis-calibrated Case	
BOT HTR	BOT PSM
15.62	5.49
16.48	6.56
15.72	6.79
14.23	5.68
17.18	8.24

The calibrated case in Table 6, exhibits an error of less than 12% of the BOT estimated by PSM with respect to the one coming from HTR measurements. This is a reasonable margin of error, given the unavoidable variations that occur in a large-scale process, as well as the intrinsic numerical uncertainty of the model when predicting the carbon and the hydrogen tails.

In Table 6, the mis-calibrated case shows a clear inconsistency in the results: the BOT calculated from the HTR results is 2–3-times higher than that estimated by the PSM. In this scenario, due to the topological constraints of the function's tail, there was no solution that can achieve alignment with the HTR case. This discrepancy strongly indicates a systematic error that affects the measurements, inflating the species yields and, consequently, producing a lower BOT value for the PSM estimation. This inflation of yields needs to be around 15% to ensure a match between the PSM and HTR calculations.

In situations where the origin of a discrepancy is unknown, potential errors can stem from various sources. These include calibration inaccuracies linked to the analytical instruments, environmental fluctuations, and inadequate observation methods. The last two will normally be perceived as random errors which increase the numerical dispersion of the data with unpredictable behavior. Contrarily, systematic errors can be perceived as consistent deviations affecting the measurements in similar way. In this case, the deliberate miscalibration caused the total volume concentration to exceed 100% for the non-condensable gases which translated into an averaged species overestimation of 10%. Such quick visible error estimation aligns well with the calculated deviation from the PSM predictions (~15%). This result underscores the ability of the PSM to pinpoint deficiencies in the quality of the experimental data. In future, PSM implementation could be further refined to assign errors specifically for each measured k-group, thereby enhancing the precision of the model's outcomes.

6.1.3. Function's Parameters Evaluation with Temperature

Figure 22 shows how the Burr parameters α_1 and α_2 vary with temperature for the three fitted distributions: carbon in olefins, carbon in paraffins, and hydrogen in olefins. Results aggregate 34 datasets from polyethylene experiments spanning multiple reactor conditions (see Table 4). The parameters were obtained via a joint optimization that simultaneously enforced experimental data fitting, overall carbon balance, and H/C-ratio constraints across the three molecular systems. The average absolute data-fit error was ~1%, and the average carbon-balance deviation was ~5% relative to the feed carbon.

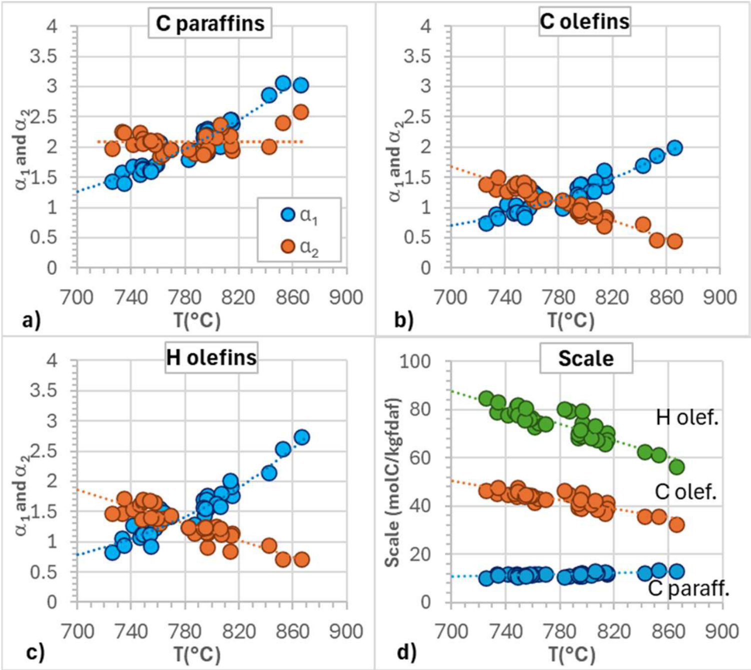


Figure 22. Burr's parameters as function of reactors temperature for paraffins' carbon (a), olefins carbon (b) and hydrogen (c), and the function´s scale (d)

From Figure 22, temperature-associated trends are observed for all three molecular systems. In every parameters panel (a to c), α_1 increases with temperature; the rise is slightly steeper for hydrogen in olefins. In contrast, α_2 decreases for carbon and hydrogen in olefins, while for carbon in paraffins it remains roughly stable in the range 2–2.5 with an average of 2.10 ($\sigma = 0.17$). All three panels show a crossing of α_1 and α_2 trends between ~760 to 790°C. The functions scale in panel d show a decreasing linear-like trend for the olefins while almost constant for the paraffins.

Based on the Burr's mathematical formulation from Table 1:

$$f_i(\alpha_1, \alpha_2) \ = \ (1 + (i-1)^{\alpha_2})^{-\alpha_1} - \ (1 + i^{\alpha_2})^{-\alpha_1}$$

The two parameters determine how the probability is allocated across the k-groups. α_1 acts as a global intensity knob: increasing α_1 steepens the overall decay, pushing probability away from higher k and concentrating mass toward the first groups; the far tail becomes thinner, and the first bin becomes taller. α_2 controls the curvature of the front end (low k-groups) by changing how fast k^{α_2} grows with k. Note that the first bin $f_1 = 1 - 2^{-\alpha_1}$ depends only on α_1, so changes in α_2 reallocate probability from the 2nd bin and forward but don't affect the first bin's magnitude. Thus, for a fixed α_1, increasing α_2 makes the distribution more front-loaded: early bins (especially i=2) gain probability, while mid bins eventually lose probability, i.e., the shape can shift from more convex to more concave at the front. Decreasing α_2 has the opposite effect: the front becomes more convex, mid bins hold more weight, and the tail becomes relatively heavier.

Therefore, at lower temperatures, longer chains are more likely to remain, and the distribution decays smoothly across k-groups. This matches the higher α_2 and lower α_1 observed below 760°C in panels a to c of Figure 22, which signals a more pronounced front-end curvature that spreads probability into mid-k while presenting a weaker overall push toward small k. As temperature rises, α_1 becomes the main driver of shape; the trends cross between 760–790 °C, suggesting a transition in fitted parameter influence from curvature-dominated to level-dominated behavior. Beyond this range, front-loading increases, consistent with higher cracking severity where the energy in the system starts to become high enough to affect most chains with similar intensity, regardless of their size, further breaking down long chains and shifting mass toward the smallest groups. At the highest temperatures, this effect leads to a pronounce skew: for olefins, probability concentrates towards ethylene ($k = 2$), and for paraffins towards methane ($k = 1$), consistent with a high-severity cracking regime (See right panel in Figure 20). Consequently, in this regime, olefins trends (panels b and c) show rising α_1 and falling α_2: the distribution becomes strongly front-loaded, more convex, and the far tail thins with a relatively stable decaying behavior as temperature increases (although the chains become more unsaturated given the steeper increase of α_1 in panel c). By contrast, paraffins maintain a relatively stable α_2 level while α_1 increases, indicating that the distribution also becomes front-loaded, but its overall decaying behavior will grow more pronounced with temperature.

From the trends observed, a simple parameter-temperature regression correlation for α_1 and α_2 can be explored as a proof of concept within the dataset window, to enable quick, approximate estimates of aliphatic molar yields from polyethylene cracking. Different kinds of fittings were tested: linear, quadratic, exponential and Arrhenius type, and the resulting R^2 values are reported in Table 7. The comparison between regression forms is primarily based on R^2 as a descriptive indicator of variance explained by temperature. Given the limited dataset size and the phenomenological nature of the regressions, R^2 is not used as a strict model selection criterion; instead, regression forms are chosen based on a balance between goodness of fit, parsimony, and physical interpretability

Table 7. Correlation coefficients for the different regression fit types performed.

		Linear		Quadratic		Exponential		Arrhenius	
M. System	**Param**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
C_pf	α1	0.90	0.13	0.91	0.13	0.91	0.13	0.91	0.13
	Scale	0.40	0.53	0.42	0.52	0.40	0.53	0.39	0.53
C_of	α1	0.84	0.12	0.86	0.11	0.86	0.11	0.86	0.11
	α2	0.92	0.08	0.92	0.08	0.89	0.09	0.87	0.10
	Scale	0.75	1.80	0.77	1.71	0.73	1.86	0.72	1.90
H_of	α1	0.85	0.16	0.88	0.15	0.88	0.15	0.88	0.15
	α2	0.81	0.12	0.81	0.12	0.78	0.12	0.77	0.13
	Scale	0.78	3.06	0.79	3.00	0.77	3.13	0.77	3.20

Consistent with the trends in Figure 22, a linear law in T is sufficient for $\alpha_2(T)$ in both olefin series. This form preserves monotonicity, avoids spurious curvature, and performs comparably to higher-order fits within the evaluated data window (see Table 7). In contrast, α_2 for paraffins (Figure 22a) exhibits minimal variation across the explored temperature range, resulting in negligible y-axis variance. In this case, a constant representation using the mean value is statistically and physically reasonable, avoiding overfitting and misleading interpretation of goodness-of-fit metrics. For $\alpha_1(T)$, non-linear regressions provide a slightly improved fit for all three systems. Although the different non-linear forms yield similar R^2 and RMSE values, an Arrhenius-type correlation is preferred due to its physical interpretability, offering a phenomenological mechanistic link between α_1 and activation-controlled severity, consistent with the fundamental temperature dependence of reaction rates and distribution characteristics. For the scale parameters, linear correlations are a suitable choice, as they provide trend-level accuracy comparable to more flexible alternatives while avoiding unnecessary complexity. Although the corresponding R^2 values are lower than for the shape parameters, they remain adequate for trend-level estimation within the evaluated range. In the paraffin case, where temperature-induced variation is minimal, a linear fit is retained to capture the modest systematic trend while acknowledging that the effect is small. The final selected parameter–temperature correlations are summarized in Table 8.

Table 8. Summary of defined parameters regression with temperature

M. System	Param	Form	a	b	R^2	RMSE
C_paraff	α_1	$\ln \alpha_1 = a + b/T_K$	6.5588	-6189.89	0.90	0.07
	α_2	$\alpha_2=\bar{\alpha}_2=a$	2.0953	-	-	-
	Scale	$s = a + b^*T_{°C}$	1.8965	0.01255	0.40	0.54
C_olef	α_1	$\ln \alpha_1 = a + b/T_K$	6.8358	-7042.69	0.83	0.1
	α_2	$\alpha_2 = a + b^*T_{°C}$	6.8964	-0.00746	0.92	0.08
	Scale	$s = a + b^*T_{°C}$	113.08	-0.09047	0.75	1.86
H_olef	α_1	$\ln \alpha_1 = a + b/T_K$	8.2166	-8279.91	0.85	0.11
	α_2	$\alpha_2 = a + b^*T_{°C}$	6.7450	-0.00698	0.81	0.12
	Scale	$s = a + b^*T_{°C}$	207.61	-0.17131	0.78	3.16

The predictive capability of these quick parameter–temperature correlations was evaluated by comparing the modeled molar yields of paraffins and olefins (k = 1 to 4) against experimental measurements. The Mean Absolute Error (MAE) and relative MAE (Re-MAE) are presented in Table 9.

Table 9. Overall parameter–temperature correlations performance against the experimental data (% yield/feed indicates elemental mole yield vs. the elemental moles of the feed in percentage)

	C Parafins				C Olefins				H Olefins			
System	k=1	k=2	k=3	Avg.	k=2	k=3	k=4	Avg.	k=2	k=3	k=4	Avg.
MAE (% yield/feed)	0.60	0.24	0.27	0.37	1.21	0.98	1.57	1.23	1.35	1.11	1.65	1.39
Rel-MAE (%)	5.10	7.12	32.44	14.88	3.96	12.57	24.65	13.73	3.60	12.30	21.98	12.61

For paraffins and olefins, the model achieved low errors for the dominant species, with Rel-MAE values of less than 12.6% for the first two predicted bins (k=1, 2 for paraffins and k=2, 3 for olefins) which indicates good agreement in the most abundant fractions. Accuracy decreased for the longer chains (k = 3 for paraffins and k=4 for olefins), where Rel-MAE reached 32.4% for paraffins and ~24.7% for olefins. However, the overall absolute yield difference given by MAE corresponds to 0.27% and less than 1.65% of the total elemental feed moles for paraffins and olefins respectively which makes no significant impact on the overall elemental balance in this dataset. In addition, the dataset used to implement the PSM included experiments under varying conditions, not only temperature but also on other factors such as bed material circulation and fuel flow, introducing natural variability that makes even such relatively large Rel-MAE discrepancy less critical in practical terms. Thus, it is possible to observe that the model presents acceptable reliability for major species that dominate the carbon balance. For larger species chains the predictions can still be considered indicative and useful as a reference, especially because these heavier species often require specialized analytical techniques for accurate quantification, which are not commonly available in all experimental setups.

Overall, it is worth noting that although visual trends exist, the observed R^2 values and data dispersion indicate that temperature alone cannot fully explain the variability. Other operational factors, such as steam-to-fuel ratio, bed material circulation and properties, and residence time, can also influence parameter behavior. Also, the selected regression forms are phenomenological and intended for quick estimation; extrapolation beyond the studied temperature range or operating regimes, should be validated with additional data. Nonetheless, despite the simplicity of the Arrhenius and linear correlations evaluated in this proof of concept, they provide reasonable trend-level accuracy for practical estimation of relevant species yields, demonstrating the potential of this approach for quick model application. It should be noted that the regression models are not intended to replace the constrained optimization; rather, they can provide a compact surrogate learned from physically constrained solutions, enabling rapid estimation while remaining aligned with physically admissible behavior within the sampled regime. Incorporating additional descriptors can be the subject of future work to enhance predictive capability while keeping the parametrization compact.

In general, the results demonstrate the potential of the PSM model to reduce the complexity of measured systems by transforming an extensive list of species into just two parameters that define a specific distribution function. This reduction in degrees of freedom allows the PSM to efficiently capture and compress information while embedding it within physics-informed frameworks, thereby adding predictive capabilities and enabling valuable insights. The findings also highlight the versatility of the PSM model as a tool for data quality assessment and ensuring the statistical and chemical coherence of results obtained from the steam cracking process, with practical applications for downstream operations and refinery control. Further details about this will be discussed in Section 7.2.

6.2. CBG model implementation for heterogenous systems

This section evaluates the CBG model for correlating heterogeneous feedstock compositions with product distributions in steam cracking. The analysis builds upon the structural classification framework introduced in **Paper II**, further explored in **Paper V**, and formalized into a predictive, condition-dependent formulation in **Paper VI**. The focus here is on the model's integration with machine-learning techniques to estimate and evaluate the mapping between feedstock bond-group fractions and product-group outputs within the CBG's low-dimensional representation. The first part examines the model under constant-temperature conditions including datasets from the mixed polymeric feedstocks studied in **Papers II**, **IV**, and **V** with varying polyolefinic shares. This evaluation tests how accurately the learned CBG-based mapping between feedstock and product simplex spaces reproduces experimentally observed product-group distributions, establishing a baseline for model structure and predictive consistency. The analysis then extends to variable-temperature conditions, assessing the model's ability to capture temperature-driven shifts in product formation within the simplified bond-group framework, as proposed in **Paper VI**. This assessment serves as a proof of concept for the condition-dependent generalization of the model. Overall, the results demonstrate the viability of advancing the CBG formulation toward predictive, physics-informed tools and reduced-order digital twin concepts for process control and optimization.

6.2.1. Heterogenous Feedstocks Evaluation and Steam Cracking Behavior

Figure 23 displays the CBG transformation of the polymeric feedstocks from Table 4 based on their polymeric composition of Figure 19 and their respective carbon groups shares (see Figure 9).

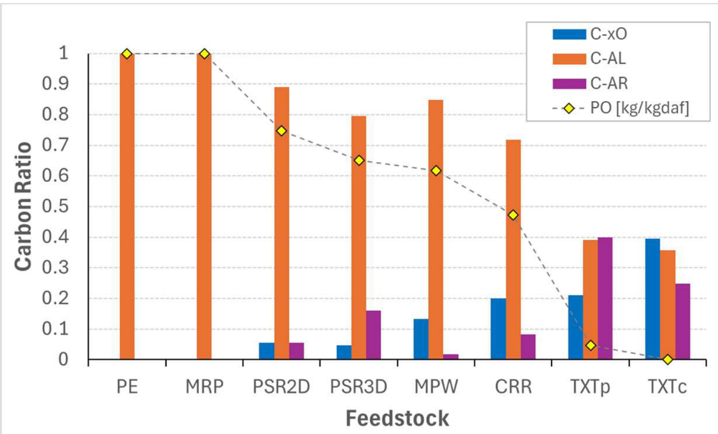


Figure 23. Carbon bond groups composition for the evaluated feedstock.(PO indicates polyolefin content).

The set of evaluated feedstocks exhibit a good variability in terms of the carbon groups, considering that all of them correspond to real waste streams. Feedstocks with large polyolefin content present the largest share of aliphatic bonds. As they incorporate other kinds of polymers, the carbon bonds are redistributed towards C-AR or C-xO depending mostly on the share of PET or cellulose in the mixture. Notice that even though the polyolefin content correlates with C-AL, non-polyolefin polymers can still contain aliphatic bonds on their structure, bringing still significant levels of C-AL even in low polyolefin feedstocks.

The results obtained at different temperature conditions for the evaluated feedstocks are presented in Figure 24.

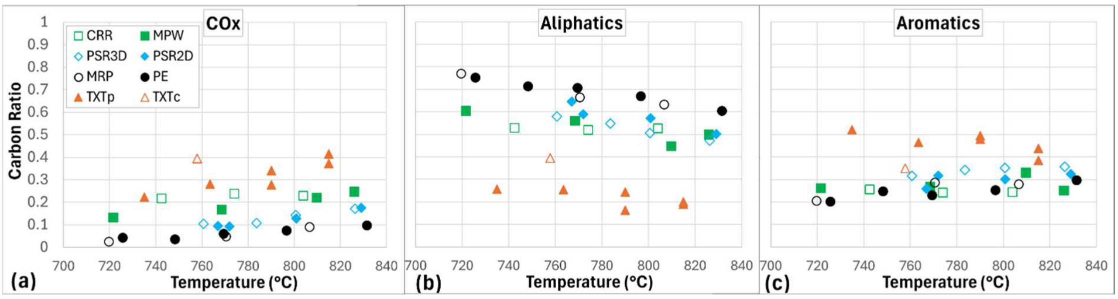


Figure 24. Carbon ratio yields of the different product groups obtained from the evaluated feedstocks at different temperature conditions.

As described in Section 4.3, the group classifications over the feed and products transform both sets into composition vectors within corresponding ternary simplex spaces. The steam cracking process is reformulated as a map between both spaces which can be modelled by a matrix link $\boldsymbol{K}$ of regression coefficients. Equations (23) and (24), sets the formulation for the CBG model considering a fixed operative regime case. Based on that, the resulting PICN optimization problem from Equation (30) and (31) was solved here with Python. The algorithm implements the first-order solver Splitting Conic Solver (SCS) from the domain-specific language CVXPY, which can be used freely as a Python library under the Apache License [77,78].

To implement the fixed-regime CBG model, the points within the relatively narrow temperature band in the middle range of Figure 24, between 760°C and 780°C, were treated as having a similar temperature, with an average of ~768 °C. Figure 25 displays the optimization results in the products ternary simplex, comparing the measured vectors with the predicted ones for the fixed-regime case.

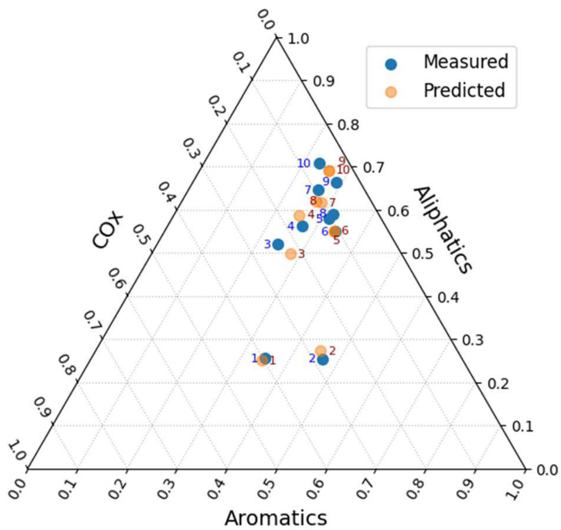


Figure 25. Measured and predicted product vector yields (Numbers in blue and red indicate point label)

For COX, Aliphatics and Aromatics the model's R^2 is respectively: [0.991, 0.977, 0.839], the mean absolute error (MAE) is: [0.0084, 0.0206, 0.0223], and the relative MAE (%): [7.09, 3.97, 7.94]

As shown in Figure 25, the model captures the overall variability of the experimental data well. The R^2 are significantly high for COx and aliphatics indicating the model explains ~99% and ~98% of the variance in these groups respectively. This suggests that their formation is well represented by the input variables under the linear transformation defined between the bonds groups and the product groups spaces. The performance for aromatics is somewhat lower at R^2~0.84, indicating that about 16% of the variance remains unexplained. The relative MAE is around 4% for aliphatics, indicating relatively good precision, and about 7-8% for COX and aromatics, which is less tight but still acceptable. In absolute terms, the MAE corresponds to prediction errors below 2.3% of the system's total carbon. Considering the scale of the reactor and the heterogeneity of the evaluated feedstocks, this indicates that the model predictions remain consistently close to the measured values, even for the aromatics group where R^2 is lower.

Figure 26 shows the cross-plots of the measured yields vectors versus the predicted ones.

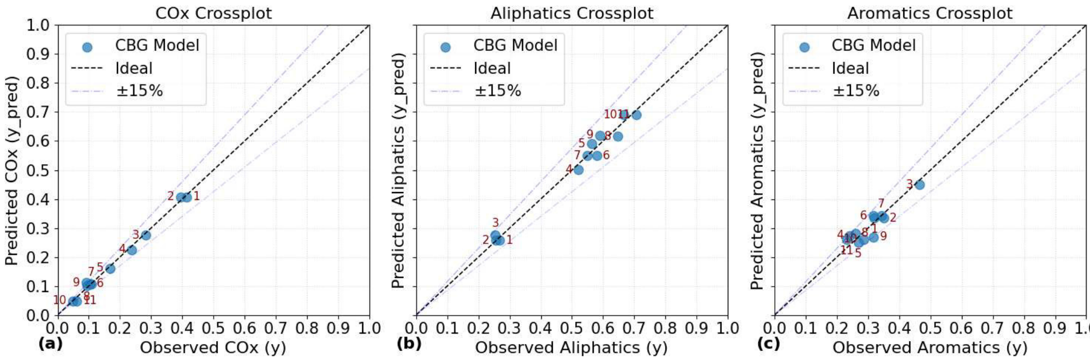


Figure 26. Cross-plots of predicted versus measured y (Numbers in red indicate point label)

As shown in Figure 26, the model predictions agrees within a good margin with the experimental measurements. All points in COx and aliphatics are very close to the diagonal line (ideal agreement), which is consistent with the high R^2 values observed for these two groups. In the case of aromatics there is relatively greater dispersion; however, most of the point remain within 15% of deviation, which is reasonably good given the conditions and the low MAE obtained.

To gain further insight into the underlying mapping, Figure 27 shows the diagonal terms of matrix $\boldsymbol{K}$ (Equation (24)) together with the results of the fitted model responses.

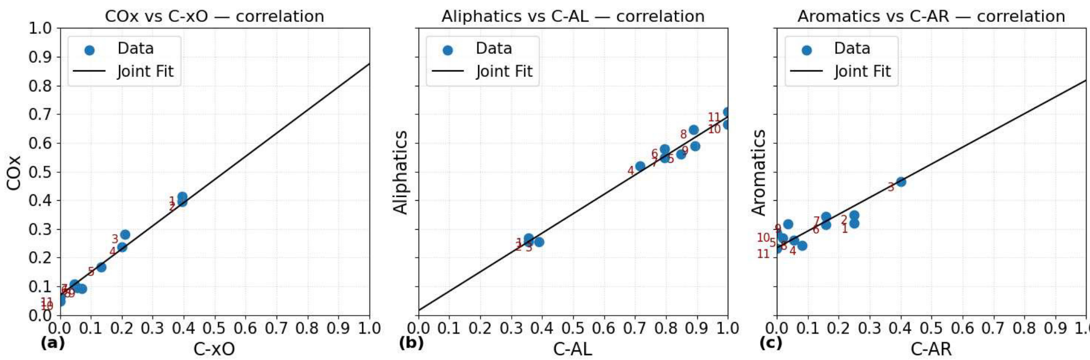


Figure 27. Diagonal pairing plots of matrix K from the CBG model at an average temperature of ~768°C (numbers in red indicate point label).

It is important to clarify that the fitted line in each panel, does not correspond to an independent ordinary least squares regression, but to the joint constrained fit model obtained for the respective linear predictor y_g (see Equation (32)). This joint constrained estimation ensures that the displayed line corresponds to the marginal representation of the physically meaningful mapping $\boldsymbol{K}$ identified for the system.

From Figure 27, clear trends can be observed for the diagonal terms. Overall, the linear fit captures with good approximation degree the system's tendency for products to remain within their corresponding chemical family. This result further supports that the fixed-regime linear mapping defined in Equation (23) between the feed and product simplex spaces provides a reasonable approximation of the system's behavior.

The non-zero intercepts obtained for panel (a) and (c) arise from contributions to those product groups from other bond classes through the off-diagonal terms of matrix $\boldsymbol{K}$. In the case of COx, this behavior is both observed and physically expected, since even a pure C-AL feed such as polyolefins produces some COx in the DFB reactor due to steam reforming and bed redox effects. For panel (c), the non-zero intercept is also expected as a consequence of cyclization reactions occurring from radical recombination reactions, which make the aromatic species immanent in the product species slate for all feed types (see Section 3.2).

The aromatics in panel (c) show some dispersion at low C-AR levels; however, the fitted line still captures the overall trend of the data reasonably well. The relatively lower R^2 observed for this group can be associated with this dispersion and may indicate the presence of more complex formation mechanisms. Such mechanisms could involve interactions with additional chemical features or operating conditions that are not explicitly represented in the model evaluated here. It should be noted that the proof-of-concept model presented in this work generalizes the chemical structure of the feedstocks and does not explicitly distinguish variations arising from the presence of branched aliphatic structures or from heteroatoms other than oxygen. Therefore, such missing features may be related to cyclization reactions involving branched methyl radicals derived from polyolefinic polymers such as PP, as well as the presence of other heteroatoms such as chloride which has a demonstrated influence at the secondary reactions level to promote aromatization in the product's slate [62,63]. These observations suggest routes for further model refinement. Although such extensions are beyond the scope of the present work, they will be briefly discussed in the future research section (Section 7).

From the optimization problem solution, the estimated K matrix for this system is given by Equation (38):

$$\begin{array}{cc} & \begin{array}{ccc} C-xO & C-AL & C-AR \end{array} \\ \boldsymbol{K} = & \begin{bmatrix} 0.8621 & 0.0495 & 0.1858 \\ 0.0075 & 0.6916 & 0.0049 \\ 0.1304 & 0.2589 & 0.8094 \end{bmatrix} \begin{array}{l} COx \\ Aliph \\ Arom \end{array} \end{array} \tag{38}$$

Recall from Section 4.3.2 that $\boldsymbol{K}$ is a column-stochastic partition map from the bond groups vector to product groups vector. Therefore, the columns show how the carbon from bond group b is expected to partition across COx/Aliphatics/Aromatics at the specific conditions. Each of the K_{gb} entries can thus be seen as conversion indices between the structural bond b in the feed to the product group g (**Paper V**), representing absorption probabilities that summarize the net transfers across many sequences of transient states in the corresponding absorbing Markov chain (Section 4.3.2).

According to the values in Equation (38), the model outcomes show a dominant diagonal structure, indicating a strong system propensity to make the bonds stay within the similar chemical product family. This behavior reflects a structural persistence of carbon environments during cracking, where many reaction pathways may conserve the underlying bond character before final product formation. The diagonal dominance is strongest for C-xO and C-AR, with conversion indices above 80% toward COx and aromatics, respectively. This is consistent with the stability of aromatic ring structures and with the tendency of oxygen-containing bonds to ultimately form energetically stable COx species. The C-AL group presents a lower diagonal conversion index of approximately 70%. This is largely due to its involvement in cyclization pathways that produce aromatic species, as reflected by the corresponding off-diagonal term, which reaches nearly 26%.

From the off-diagonals, propensity to reforming or gasification pathways is significantly higher for the C-AR group (18.6%) than for C-AL (5%). In contrast, the C-AL group exhibits a stronger involvement in aromatization reactions, as stated previously. A noticeable conversion from C-xO to aromatics (~13%) is also observed. This behavior may be related to deoxygenation pathways in oxygen-containing structures that generate unsaturated intermediates capable of participating in secondary cyclization and aromatization reactions. Negligible transfers are observed from C-xO and C-AR to aliphatics, with the latter even falling below the imposed domain constraint $\epsilon = 0.01$ (see Equation (31)), which gives indications that the data supports such chemistry prior.

Overall, these observations illustrate how the estimated matrix Ksummarizes the dominant carbon-routing pathways within the reactor. The results represent the averaged behavior of the modelled cracking system for the evaluated feedstocks under relative similar temperature levels. The next section presents the results of the conditions-dependent CBG model, where the system is analyzed under a variable temperature regime.

6.2.2. Conditions-dependent CBG Model Implementation

Figure 28 shows the results of the generalized CBG model for operative conditions of Equation (33) with the full set of experimental cases of Table 4, see Figure 24 (N=32). As mentioned in Section 4.3.4, temperature was taken as the only driving condition for $\phi(u)$ in a linear relationship for the proof of concept. To get good performance, the solution algorithm was based on the ADAM gradient descent method and implemented via the torch library in Python along with tensorial operations at the core of the calculations with the different arrays. Training was done with over one thousand epochs[1] to get a stable minimal loss level with a learning rate of 0.05 and other ADAM hyperparameters by default ($\beta_1 = 0.9, \beta_2 = 0.999$) [72].

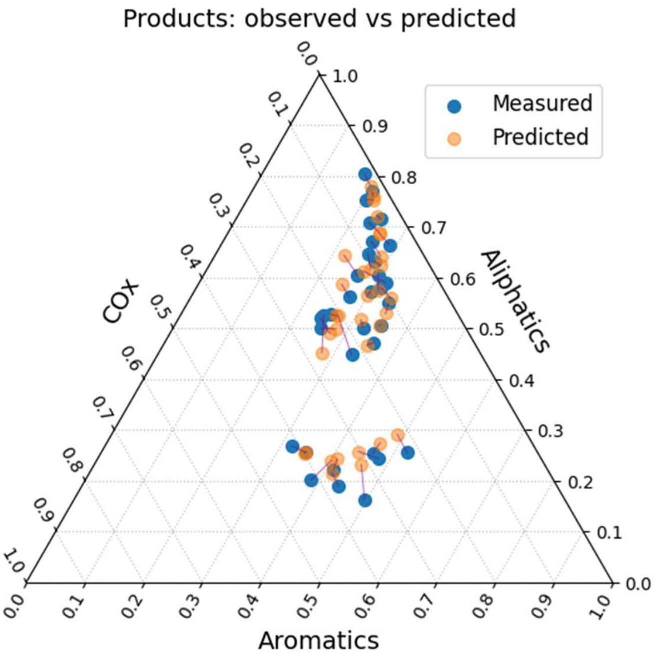


Figure 28. Results of the conditions-dependent CBG model from the full dataset of experimental cases (small purple lines connect the paired points between measured and predicted).

[1] An epoch corresponds to a complete pass of the whole dataset to train the model. During each epoch, the model processes all the training samples, updating its parameters to learn the patterns and conditions that minimize the defined loss function.

For COX, Aliphatics and Aromatics the R^2 of the model is respectively: [0.977, 0.972, 0.903], the mean absolute error (MAE) is: [0.0126, 0.0230, 0.0218], and the relative MAE (%): [8.19, 6.44, 7.17].

As it is possible to observe from Figure 28, the generalized CBG model displays a notoriously good correlation with the variability of the measured data. In general, the predicted vectors follow satisfactorily well the measured ones in the evaluated space, which imply the model is flexible enough to capture the behaviors behind the obtained yields for the different feedstocks and temperature conditions evaluated. The R^2 is still at a good level with the model explaining around 97% of the variance in the COx and aliphatics group yields. For aromatics, the R^2 increased to 0.90 when compared with the averaged-temperature case of previous section. Although there is still ~10% of the aromatics variance the model needs to account for, its performance increased slightly. This outcome might come in a large part from the incorporation of new data into the model. However, the consideration of temperature variations in the formulation through the nonlinear SoftMax link can definitely play a role in improving the model's performance, especially since the aromatics formation is sensitive to the system's temperature level.

Despite that relative MAE increased slightly for COx and aliphatics, which might be in part due to the presence of some outliers (as will be discussed later), the model's MAE is still below or kept at around 2.3% of the carbon for the three groups, which is reasonably well considering the variable conditions regime that now has been implemented in the model.

<u>*Cross-plots Analysis*</u>

Figure 29 presents the cross-plots comparing measured and predicted product yields for the condition-dependent CBG model.

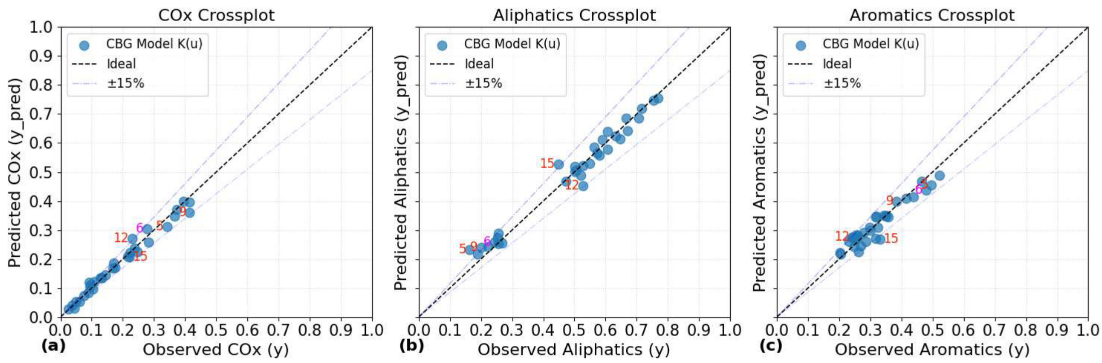


Figure 29. Cross-plots of the predicted vs the measured product yields for the conditions-dependent CBG model (colored numbers indicate point's label)

Overall, the results in Figure 29 reveal that the model provides a good approximation across all three product groups, capturing the variability of the data over the full yield span within reasonable deviation bounds. The red-labeled points indicate cases approaching the ±15% deviation limits.

A closer inspection shows that points 9, 12, and 15 correspond to the highest temperature cases for TXTp, CRR, and MPW feedstocks, respectively, while point 5 corresponds to a mid-to-high temperature TXTp case. With the exception of point 5, the largest deviations correspond to extreme experimental cases. For instance, point 15 shows overprediction in aliphatics and a comparable underprediction in aromatics, suggesting a redistribution effect within the constrained mapping. Points 9 and 12 display a similar pattern for COx and aliphatics.

At this stage, the model is significantly constrained by the available dataset and the imposed domain priors. Consequently, deviations at extreme operating conditions may reflect either limited representativeness of the training data in those regimes or additional process sensitivities not fully captured by the current reduced-order formulation. In the case of point 5, comparison with point 6 (colored magenta), which has similar feedstock type and temperature but significantly smaller deviations, suggests that localized experimental variability may also contribute to the observed discrepancy.

It is worth mentioning that some of these cases also differ in operating parameters such as bed material circulation rate, feed mass flow, and feed batch composition, which are not explicitly represented in the current model and may introduce additional variability. Nevertheless, the overall agreement across these heterogeneous conditions, within acceptable deviation limits, supports the predictive robustness of the model when applied to mixed datasets from a semi-industrial reactor.

Beyond predictive performance, these observations highlight the model's potential as a quick data-quality assessment tool to flag conditions and regimes where additional data or refined modelling assumptions are required. Expanding the dataset in these regions can not only serve to improve the model's robustness and performance metrics, but also to validate the consistency of measurements against possible data outliers.

Generalization Error Test

To assess the model's generalization error, i.e., the robustness of the model against unseen data, k-fold was chosen as a cross-validation technique. This is a widely used method that repeatedly split the data to test the average model's ability to generalize and is especially suitable to evaluate models with small to medium sized datasets. Table 10 shows the results for 5 folds and also the result for leave-on-out (LOO)[1]:

Table 10. K-fold and LOO cross-validation results

	R^2			MAE			Relative MAE (%)		
Fold	**COx**	**Aliph**	**Arom**	**COx**	**Aliph**	**Arom**	**COx**	**Aliph**	**Arom**
1	0.909	0.931	0.829	0.018	0.028	0.028	12.23	5.49	9.88
2	0.980	0.971	0.855	0.011	0.024	0.025	8.13	6.20	7.47
3	0.946	0.989	0.737	0.030	0.018	0.026	13.14	5.55	8.62
4	0.973	0.974	0.895	0.014	0.026	0.027	8.50	9.98	7.49
5	0.972	0.942	0.803	0.013	0.029	0.030	9.67	7.56	10.49
Avg.	**0.956**	**0.961**	**0.824**	**0.017**	**0.025**	**0.027**	**10.34**	**6.95**	**8.79**
LOO Avg.	**-**	**-**	**-**	**0.015**	**0.025**	**0.025**	**9.70**	**6.79**	**8.26**

According to the averaged results of Table 10, the model demonstrates good performance across both k-fold and LOO cross-validation, indicating stability and predictive reliability. The R^2 values show that the model still captures most of the variance in the data, especially for COx and Aliphatics, indicating its robustness even when trained with a relatively small sample set.

[1] The k-fold cross-validation partitions the dataset into k equal-sized subsets (folds). The model is trained and validated k times (usually 5 or 10), each time using a different fold as the validation set and the remaining folds as the training set. LOO is the special case when k=N_{samp}, each fold contains exactly one sample for validation, and the rest for training. It gives an unbiased estimate of generalization error especially for small datasets.

MAE and relative MAE values are consistently low across both validation schemes, with LOO yielding marginally better scores due to its larger training sets per fold. This consistency implies that the actual state of the model is not overly sensitive to specific data partitions and can generalize well to unseen samples. Still the model can increase its robustness by the acquisition of more data, to further reduce R^2 generalization error especially in the case for aromatics.

The low error margins reinforce the model's utility for experimental prediction and design, especially in optimizing feedstock composition or process conditions. Overall, the model balances accuracy and robustness, making it a good performing tool for interpreting and forecasting the evaluated product groups in the steam cracking system.

K Matrix Temperature dependence

Figure 30 shows the variation of the matrix ***K*** entries as function of the evaluated temperature conditions.

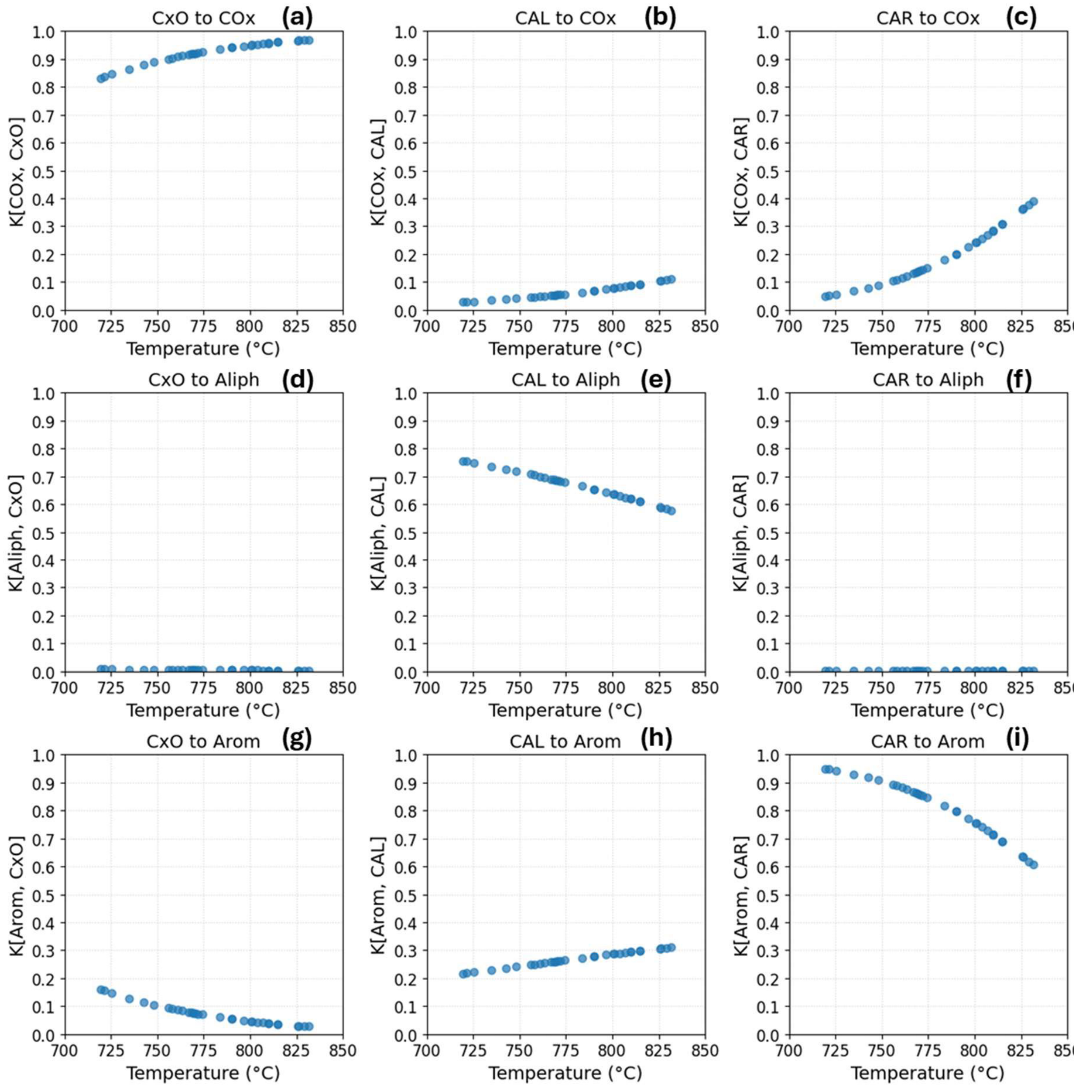


Figure 30. Matrix ***K*** entries variation as function of temperature from the conditions-dependent CBG Model.

As observed in Figure 30, the conversion indices present different monotonic trends as temperature increases, depending on the type of bond-to-product transfer pathway. The transference from C-xO to COx in panel (a) starts at an index around 0.83 at the low temperature case and displays a non-linear behavior towards one as temperature increases, while the corresponding C-xO to Arom in panel (g), decays in similar fashion towards zero starting from ~0.16. C-AR to Arom in panel (i), exhibits a steeper non-linear decaying behavior starting from ~0.95 and ending up at ~0.61 as temperature increases in the same span. In contrast, the transference from C-AR to COx in panel (c) increases in alike steeper behavior starting from ~0.05 till ~0.39. C-AL to aliphatics (panel(e)) displays an almost linear decaying behavior from ~0.75 to ~0.58, whereas C-AL to COx (panel(b)) increases non-linearly from 0.03 to 0.11. C-AL to Arom in panel (h), increases with mild non-linear behavior from ~0.22 to ~0.31.

Therefore, according to these model's outcomes, the likelihood of a C-xO bond type to end up as COx species increases as temperature increases. This is indeed a plausible behavior. Higher thermal energy in the system favors breaking down; thus, simpler and more stable structural forms such as COx will prevail from newly released oxygen-containing fragments. With such chemical preference, all other species' formation paths must tend to decrease in turn.

In the case of C-AR to Arom (panel i), the large index at low temperatures reflects the intrinsic stability of the aromatic ring which will tend to conserve its structural configuration. As temperature increases, ring-opening steam-involving reactions become favored, increasingly affecting the aromatics structures and leading to the production of COx. The model thus captures a redistribution from structural preservation toward oxidative conversion which aligns with the fact that gasification regime starts to gain momentum as temperature increases, until eventually reaching a level where all aromatic structures may become involved in such conversion.

Negligible conversion indices for C-xO and C-AR to aliphatics are obtained, consistent with a low likelihood of forming stable aliphatic species directly from oxygenated or aromatic carbons under the studied conditions. This also relate to the apparent mirrored-like behavior in the respective Figure's corner plots. Relaxation of the constraint imposed to aliphatics from C-AR in Equation (31) did not alter this result, suggesting that the constraint is not binding as per the actual conditions of the system; so, the information is carried out by the data itself. An appreciable change might be expected in the index in case a catalytic effect in the system starts to break aromatic ring radicals to yield aliphatic species.

In the case of C-AL, the behavior at low-temperature is dominated by retention in aliphatics. Off-diagonal conversions, present low contribution to COx at this regime and a main focus in cyclization reactions to produce aromatic species, as it can be corroborated by the yields of COx and aromatics in Figure 24 for pure C-AL feeds such as PE and MRP. As temperature increases, conversion to COx starts to ramp up non-linearly. The conversion to aromatics continues increasing but following a mild concave down trend. This eventually may achieve a maximum level at some higher temperature while reforming reactions will continue to take place with an increasing relevance in the conversion process.

Final Remarks

An additional implication of these results lies in the interpretation of the matrix $\boldsymbol{K}$ as the transformation between the feed and product simplex spaces, effectively representing the reactor at the evaluated level of abstraction. The set of $\boldsymbol{K}$ entries and their behavior with temperature relates thus to the averaged performance of the reactor at such dimensional-reduced order. Therefore, the better the model ‘s performance in capturing the conversion’s behavior at different conditions and feeds, the more suitable will $\boldsymbol{K}(T)$ become a digital twin of the reactor (Figure 31).

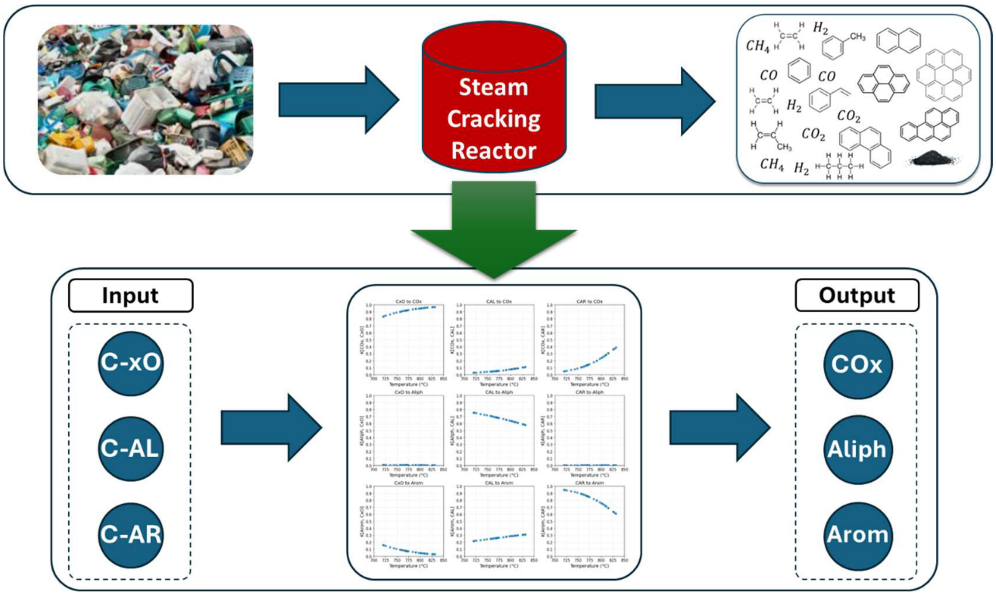


Figure 31. K matrix as a digital twin of steam cracking reactor at an abstracted level.

Once a validated and generalizable $\boldsymbol{K}$ mapping is established for a given reactor configuration, it can enable rapid estimations of product distributions within its operating domain. Such a representation can support real-time monitoring and decision-making by providing physically constrained predictions at negligible computational cost.

Overall, these results constitute a proof of concept for a generalized CBG-based reactor mapping. Future developments may extend the dimensionality of the framework in terms of structural descriptors and product basis to capture additional variance, incorporate further operational variables, and integrate the framework with complementary models to enrich feed characterization and product resolution. These perspectives are discussed in Section 7.

7. Future Perspectives

7.1. Generalized Reactor-Dependent CBG model and Carbon Groups Decoupling Extension

7.1.1. Generalized reactor-dependent CBG model

The CBG framework developed in Section 4.3 represents the cracking process as a mapping between the feed bond groups and the observed product distribution. This mapping is expressed in an extended form through an absorption stochastic matrix $\boldsymbol{K}(\boldsymbol{u})$ that depends on a set of operator-controlled conditions $\boldsymbol{u}$ (temperature, steam dilution, etc.). In this work, the transformation was treated as a single, condition-dependent model. However, the reactor's configuration itself may also exert a relevant influence on the observed yields. Factors such as heat transfer efficiency, mixing behavior, temperature profile, ash accumulation, wall effects, among others, introduce systematic biases that are mostly intrinsic to the reactor and cannot be easily measured or manipulated as freely as an operative variable.

Such factors may influence certain product tendencies which cannot be fully captured by just feedstock characteristics or operator-controlled conditions. For instance, some reactors may systematically favor the formation of aromatics or enhance syngas yields regardless of the precise feedstock. To model these effects, a pseudo-bond channel (ps) can be introduced. This pseudo-bond does not represent an actual type of bond but rather a reactor fingerprint that accounts for background conversion pathways. Such pseudo-bond artifact can be applied to the $\boldsymbol{K}$ matrix and the vector composition $\boldsymbol{f}$ as a mathematical strategy to get the model's enhancement that considers such systematic effect. Equation (33) is then transformed as:

$$\boldsymbol{y}(\boldsymbol{u}) = \widetilde{\boldsymbol{K}}(\boldsymbol{u})\tilde{\boldsymbol{f}} \quad \Rightarrow \quad \widetilde{\boldsymbol{K}}(\boldsymbol{u}) = [\boldsymbol{K}(\boldsymbol{u}) \quad \boldsymbol{k}^{ps}(\boldsymbol{u})]; \;\; \tilde{\boldsymbol{f}} = \left[\big(1-\alpha(\boldsymbol{u})\big)\boldsymbol{f} \quad \alpha(\boldsymbol{u})\right]^{T} \tag{39}$$

Here, $\boldsymbol{K}(\boldsymbol{u})$ captures what can be explained from the bond structure information in $\boldsymbol{f}$; $\boldsymbol{k}^{ps}(\boldsymbol{u})$ is a column vector corresponding to a normalized product distribution that characterizes the reactor's specific fingerprint that captures such tendencies that cannot be explained by $\boldsymbol{f}$, and $\alpha(\boldsymbol{u}) \in [0,1]$ is a weight controlling its influence. Thus, the new mapping is described by:

$$\boldsymbol{y}(\boldsymbol{u}) = \big(1-\alpha(\boldsymbol{u})\big)\boldsymbol{K}(\boldsymbol{u})\boldsymbol{f} + \alpha(\boldsymbol{u})\boldsymbol{k}^{ps}(\boldsymbol{u}) \tag{40}$$

When $\alpha(\boldsymbol{u}) = 0$, the output depends purely on bond composition; when $\alpha(\boldsymbol{u}) > 0$, a fraction of the output is being shaped by the pseudo-bond channel. $\alpha(\boldsymbol{u})$ is a trainable parameter and can be defined in a first approximation as $\alpha(\boldsymbol{u}) \sim \sigma\big(\gamma^{\top}\phi(\boldsymbol{u})\big)$, using sigmoid function (σ) to keep the domain $[0,1]$. It is worth to mention that this formulation does not add mass to the system, but rather redistributes part of the compositional control between bond-driven ($\boldsymbol{K}$) and reactor-driven tendencies ($\boldsymbol{k}^{ps}$).

Both the columns of $\boldsymbol{K}(\boldsymbol{u})$ and the pseudo-bond vector $\boldsymbol{k}^{ps}(\boldsymbol{u})$ are parameterized through a SoftMax transformation of logits in a similar manner as done in Equation (34):

$$Z_{gb}(\boldsymbol{u}) = A_{gb} + \left(B_{gb}\right)^T \phi(\boldsymbol{u}), \qquad K_r(u)_{gb} = \frac{exp\left(Z_{gb}(\boldsymbol{u})\right)}{\sum_{g'} exp\left(Z_{g'b}(\boldsymbol{u})\right)} \tag{41}$$

$$\zeta_{gb}(\boldsymbol{u}) = A_{gb}^{ps} + \left(B_{gb}^{ps}\right)^T \phi(\boldsymbol{u}), \qquad k^{ps}(u)_{gb} = \frac{exp\left(\zeta_{gb}(\boldsymbol{u})\right)}{\sum_{g'} exp\left(\zeta_{g'b}(\boldsymbol{u})\right)} \tag{42}$$

Where $\phi(\boldsymbol{u})$ is the feature map of the operative conditions. The SoftMax transformation ensures that both $\boldsymbol{K}(\boldsymbol{u})$ and $\boldsymbol{k}^{ps}(\boldsymbol{u})$ remain strictly positive and column-stochastic for all values of $\boldsymbol{u}$. By applying regularization penalties to these parameters, the model remains identifiable and avoids overfitting.

The resulting reactor-dependent CBG model provides a generalized description of steam cracker systems. Through the mathematical approximation of a pseudo-bond, the model can capture product distributions dependances on effects tied to reactor configuration. This model's extension has also the potential to build up towards reactor scaling relationships, where information from lab-scale experiments can be compared or translated to semi-industrial conditions within a unified data-analysis framework.

7.1.2. Carbon Groups Decoupling Extension Consideration

As presented in Section 4.2.2, the CBG framework relies on a reduced set of bond groups and product groups that allow complex systems to be represented in a tractable way. The standard formulation distinguishes between three groups for the input and output spaces. This coarse resolution has the advantage of simplicity and stability, but it may mask important structural differences in the feed or relevant differences in the products. A decoupling strategy can therefore be considered to increase the model's resolution by introducing finer sub-groups so that it can differentiate between feeds or products that otherwise appear similar. For example, polyethylene (PE) and polypropylene (PP) have nearly identical ultimate analysis, but their molecular structures differ by the attached methyl groups in PP. Introducing a methyl-specific bond subgroup to the CBG space, would allow the model to distinguish contributions of CH_2 units from those of CH_3 groups, enabling improved discrimination between PE-rich and PP-rich mixtures. Similarly, the aliphatic group could be split further to separate carbons in double bonds (=C), which are more reactive and closely linked to olefin and aromatic formation. On the product side, refining the categories could allow aliphatic monomers, and light aromatics to be predicted individually from their respective lumped groups. Additional extensions may consider separating carbons bound to specific heteroatoms, such as Chlorine, which is known to influence cracking pathways and aromatization behavior. Such refinements have the potential to increase the model's explanatory power and process-level insight.

It is important to recognize that increasing dimensionality introduces a trade-off. Pros include richer interpretability, better discrimination between similar feeds, and potential improvement of mechanistic links (e.g., methyl groups to propylene yields). To maintain model stability, additional constraints must be imposed. Sub-groups must always sum back to their parent group (e.g., C–AL = CH_3 + CH_2 + CH + =C). Hydrogen balance constraints may become necessary, because each subgroup carries a specific H/C ratio. In the product side, aggregation operators can be defined to re-lump fine groups into the original coarse categories, ensuring backward compatibility and enabling comparison across models of different resolutions. On the other hand, as cons it may have: greater data requirements to identify the increased number of parameters in matrix $\boldsymbol{K}$, and the need for detailed product characterization of the related species. To avoid overfitting with the increased number of free parameters in $\boldsymbol{K}$, the experimental design must be adjusted. At minimum, the number of independent feedstock cases should be sufficient to span the higher-dimensional bond space. For instance, to identify the effect of methyl carbons, experiments with PP-rich feeds must be included as well as relevant mixtures compositions.

Ultimately, the balance lies in selecting subgroup refinements that add meaningful explanatory value while remaining compatible with data availability and analytical capability. As datasets expand and characterization methods improve, such extensions can be implemented incrementally, preserving the tractability of the CBG framework while progressively enhancing its descriptive power.

7.2. PSM and CBG Model Integration in Reactor's Control and Operation

7.2.1. Physics-Informed Neural Network (PINN) Integration Concept for Product Species Prediction

Effective reactor control requires fast prediction of key product yields under varying feedstocks and operating conditions. Detailed kinetic models are often too computationally demanding for real-time applications, while purely data-driven models may generate outputs that violate conservation laws or physical consistency. A practical alternative is to integrate the physics-informed structure embedded in the CBG and PSM models within a robust pure data-driven framework, such as a neural network (NN) model with a perceptron architecture capable of representing complex nonlinear relationships in the data. Such integration can be conceived as a form of physics-informed neural network (PINN).

In a first conceptual formulation, the NN learns to map from an input set $(\boldsymbol{u}, \boldsymbol{x})$, operative conditions and polymeric fractions, to outputs $\hat{y}$, the predicted product distribution. Crucially, the learning process should not be free to choose arbitrary mappings. Within a unified PINN framework, physics-based transformations must either be embedded as structured intermediate layers or enforced through constraint-based loss terms. These constraints include conservation of carbon, non-negativity of yields, normalization of distributions, admissible hydrogen balances and carbon product groups. The CBG and PSM models can provide a viable scaffold for introducing such structure. A conditions-dependent CBG model, can transform a given $\boldsymbol{x}$ composition under conditions $\boldsymbol{u}$ into the main product classes such as COx, Aliph and Arom through the mapping $\boldsymbol{K}(\boldsymbol{u})$. Once the main product distribution is determined, the PSM model adds resolution to the aliphatic fraction through compact parametric distributions for paraffins and for olefins. This layer will enable the prediction of detailed key yields in aliphatics, which are often the primary control targets in industrial cracking. Overall, by incorporating these structured transformations, the network is expected to remain anchored to physically admissible behavior while retaining flexibility to adapt to data (Figure 32).

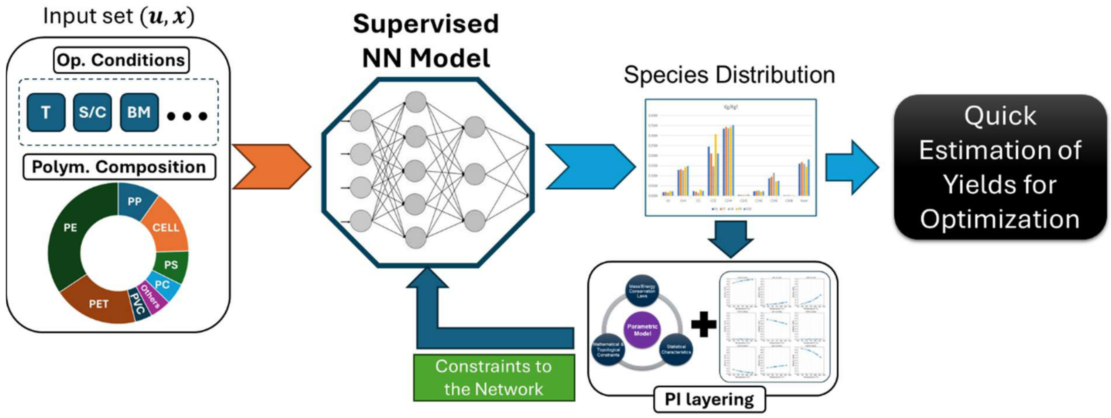


Figure 32. Schematic of PINN integration concept with PSM and CBG models

Although training the PINN will require a good enough pool of experimental data covering different feedstocks and operating conditions. Not every species needs to be measured in every sample: the network can handle partial information by only penalizing errors in the outputs that are available for a given experiment. Over time, the model integrates information across datasets, filling gaps and building a physically consistent representation of the process. Once trained and validated, the model can deliver quick and meaningful predictions of product distributions for any new input set within the trained domain.

Although this is a conceptual idea that requires further methodological refinement for its implementation, the potential advantages of an integrated modelling approach can be several. First, it enables fast evaluation based on a set $(\boldsymbol{u}, \boldsymbol{x})$, suitable for real-time control. Second, it can enforce physical admissibility, preventing the generation of unfeasible results such as negative yields or balance violations. Third, generalization across different feedstocks by the information from the pool of data and by the enforcement of the $\boldsymbol{K}$ matrix behavior. And fourth, it may be able to directly produce specific key species yields relevant to the industrial operation. Therefore, by combining physics-informed structure with data-driven flexibility, the proposed PINN framework represents a promising direction for future reactor monitoring and optimization tools.

7.2.2. Polymeric composition estimation for unknown plastic mixtures

For effective control and optimization of the cracking conversion process of heterogeneous polymeric feedstocks, knowledge at all times of the incoming feed composition is highly desirable to optimize the process towards an economically favorable species distribution. In typical plastic waste streams, however, the polymeric composition is often unknown or only partially characterized. **Paper II** considered the strategy of numerically estimate the feedstock composition based on solving a CN instance created from the elemental, mass and energy conservation via convex optimization. Using the ultimate analysis (UA) and Low Heating Value (LHV) of the Feedstock, it is possible to build up layers of constraints to estimate the polymeric mass fractions based on the same set of features from the pure polymers contained in the feed. Equation (43) presents the formulation of such system (here Einstein's notation for matrices and vectors is used for easiness of reading).

$$\boldsymbol{Cx} = \boldsymbol{V} \;\; \Rightarrow \sum_j C_j^m \, x^j = V^m; \quad \boldsymbol{C} = \begin{bmatrix} a_j^i \\ LHV_j^f \\ \boldsymbol{1}_j \end{bmatrix}, \boldsymbol{V} = \begin{bmatrix} a^{i,f} \\ LHV^f \\ m_{tot}^f \end{bmatrix} \tag{43}$$

Where a_j^i is the mass fraction matrix ($i \times j$) of the element $i \in \{C, H, O, N, Cl, \dots\}$ in polymer $j \in \{PE, PP, PVC, \dots\}$ (units: kg_i/kg_j), x^j is the mass fraction vector ($j \times 1$) of polymer j within the feedstock blend f (units: kg_j/kg_f), and $a^{i,f}$ is the mass fraction vector ($i \times 1$) of element i in the feedstock f (kg_i/kg_f).

As seen in Section 4.2.2, a feedstock mixture can also be represented in a reduced bond-group simplex space via the CBG transformation (see Figure 10). Therefore, the CBG model can constitute an additional set of equations that adds one more level of restriction to the system. Such constraint layer is of high relevance because the measured cracking products contain valuable information on the chemical structure of the feed, and the CBG model creates a link between the outputs and the inputs at a dimensional reduced level.

Then, if a validated condition-dependent mapping $\boldsymbol{K}(\boldsymbol{u})$ is available, by combining Equations (19) and (33),the additional CBG- derived layer of equations can be expressed as:

$$\boldsymbol{y}(\boldsymbol{u}) = \boldsymbol{K}(\boldsymbol{u}) \frac{\boldsymbol{Dx}}{\boldsymbol{1}^T\boldsymbol{Dx}} \Rightarrow k\boldsymbol{y}(\boldsymbol{u}) = \boldsymbol{K}(\boldsymbol{u})\boldsymbol{Dx} = G(\boldsymbol{K})\boldsymbol{x} \tag{44}$$

Where k is a constant related to the carbon share in the feed and G is a function of $\boldsymbol{K}(\boldsymbol{u})$. Introducing Equation (44) into Equation (43), the final augmented CN's system of constraints is obtained:

$$\widetilde{\boldsymbol{C}}\boldsymbol{x} = \boldsymbol{V} \Rightarrow \begin{bmatrix} a_j^i \\ LHV_j^f \\ \boldsymbol{1}_j \\ G(\boldsymbol{K}) \end{bmatrix} x^j = \begin{bmatrix} a^{i,f} \\ LHV^f \\ m_{tot}^f \\ k\boldsymbol{y}(\boldsymbol{u}) \end{bmatrix} \tag{45}$$

The resulting problem can be solved under non-negativity and convex constraints, potentially with regularization to mitigate sensitivity to possible measurement noise. This new additional layer augments the original linear system of equations based on mass and energy analysis with chemical structural information inferred from the thermochemical conversion outcomes.

7.2.3. Steam Cracker as a Structural Analyzer and Adaptive Feedback Concept Framework

Steam cracking is conceived here as the core reaction step in the plastic waste conversion process. From a conceptual standpoint, given the right setup, enough quality on the characterization of the product species, and reliable energy measurements, the process may provide sufficient information to obtain a first estimation of the main elemental composition (C, H, O ,N, S, Cl) as well as LHV of the feed. This perspective could unlock a form of online, continuous system analysis. Since the reactor operates under continuous feed and can process large feedstock volumes, it offers the potential to determine bulk, time-averaged properties of heterogeneous feeds in near real-time. The polymeric composition can be then inferred from the augmented PICN of Equation (45), using estimated UA and LHV plus the CBG-derived equations. Such estimates, along with details of the product species distribution, could be incorporated within a backwards optimization framework (using, for instance, the pre-trained PINN described previously) to provide a feedback-driven control loop to refine adaptively the reactor conditions towards optimize desired species throughputs.

Additional direct polymeric composition characterization methods such as Near Infrared Spectroscopy (NIR) and also traditional UA from batch sampling can be performed at the feedstock's preprocessing stage, to provide external estimators on the solution space that can help with the calculations speed and validation.

Figure 33 illustrates the model-guided operative conditions refinement conceptualization. While further methodological development and validation are required, this framework represents a promising direction for integrating structural modeling, process measurements, and optimization in heterogeneous waste steam cracking systems.

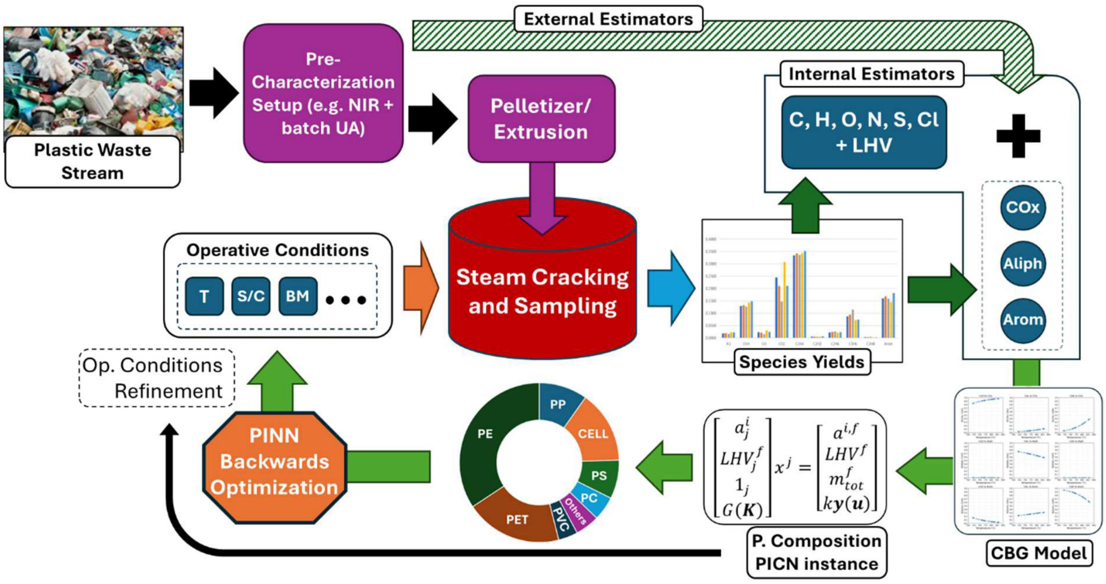


Figure 33. Schematic of loop process control for an adaptative, model-guided operative conditions refinement concept.

8. Summary and Conclusions

This thesis presented a constraint-centric methodology for analyzing and modelling steam cracking of polymer-rich feeds in dual fluidized bed (DFB) systems. The central premise is that high-dimensional, partially observed, and heterogeneous experimental thermochemical data can be rendered reliable and actionable when fused with first-principles constraints and chemically interpretable state reductions. To this end, two complementary modelling constructs were developed and applied to a pool of experimental data obtained from processing polymer-based materials in a semi-industrial scale DFB steam cracker reactor: (i) the Parametric System Model (PSM), a physics-informed parametric representation of product distributions embedded in a Constraint Network (CN); and (ii) the Carbon-Bond-Group (CBG) framework, which maps feedstock structure to product groups through transformations between compact structure-based simplex spaces. Key contributions can be highlighted from the present work:

- Formalization of CN frameworks for thermochemical conversion process data. The work encoded systems invariants (such as mass and elemental balances), carbon-class scope consistency, and topological/shape constraints as feasibility sets and features that bound estimation. This structured framework enabled the development of PICN-based models for steam cracking capable of delivering descriptive, predictive, and prescriptive analytics grounded in first principles.
- Development of the PSM. The PSM transformed structure-indexed product groups into low-dimensional parametric distributions consistent with conservation and shape constraints. By assembling a PICN with conservation laws and statistical/topological conditions the model yielded closed-form, low-parameter solutions that provided meaningful predictions of unmeasured species and data-quality validation for experimental assessment.
- Introduction of the CBG framework. The CBG model reduced both feed and product spaces to interpretable, low-dimensional vectors based on bond environments and product families. Steam cracking conversion was represented as a column-stochastic mapping from feed carbon-bond classes to product groups, enabling cross-feed comparisons on a common structural basis and systematic evaluation of product variations arising from feed composition and operating conditions. The condition-dependent mapping further revealed chemically interpretable carbon-routing trends across temperature regimes.
- By casting reactor behavior into constrained, low-dimensional representations, the resulting models have the potential to be used for product distribution estimation, reconciliation, and supervisory control, transforming complex thermochemical data into actionable feedback.

The results demonstrate that embedding conservation laws, chemically meaningful dimensionality reduction, and conditions-aware mappings into a coherent CN framework yields models that are predictive, interpretable, and auditable. The complementary PSM and CBG constructs together provide a structured basis for developing actionable frameworks for process design and operation. Beyond DFB steam cracking, the CN-based methodology is generic to conversion systems that admit a consistent carbon-class (or analogous) basis. By aligning data analytics with first principles, this work establishes a methodological foundation for models that can be useful in real-time monitoring, optimization, and adaptive control of plastic waste thermochemical recycling processes, such as DFB steam cracking advancing further its integration into circular plastics value chains.

REFERENCES

[1] M. Dokl, A. Copot, D. Krajnc, Y. Van Fan, A. Vujanović, K.B. Aviso, R.R. Tan, Z. Kravanja, L. Čuček, Global projections of plastic use, end-of-life fate and potential changes in consumption, reduction, recycling and replacement with bioplastics to 2050, Sustain. Prod. Consum. 51 (2024) 498–518. https://doi.org/10.1016/j.spc.2024.09.025.

[2] K. Houssini, J. Li, Q. Tan, Complexities of the global plastics supply chain revealed in a trade-linked material flow analysis, Commun. Earth Environ. 6 (2025) 257. https://doi.org/10.1038/s43247-025-02169-5.

[3] OECD, Lifecycle of plastic waste worldwide in 2019 (in million metric tons)[graph], (2019).

[4] K. Ragaert, L. Delva, K. Van Geem, Mechanical and chemical recycling of solid plastic waste, Waste Management 69 (2017) 24–58. https://doi.org/10.1016/j.wasman.2017.07.044.

[5] E. Anglou, F. Boukouvala, P.M. Stathatou, A cradle-to-gate life cycle assessment of mechanochemical recycling of poly(ethylene terephthalate) waste, Chemical Engineering Journal 524 (2025) 168719. https://doi.org/10.1016/j.cej.2025.168719.

[6] F. Kähler, M. Carus, C. vom Berg, M. Stratmann, CO_2 Reduction Potential of the Chemical Industry Through CCU, 2022. https://doi.org/10.52548/UTRL5869.

[7] M. Rosental, T. Fröhlich, A. Liebich, Life Cycle Assessment of Carbon Capture and Utilization for the Production of Large Volume Organic Chemicals, Frontiers in Climate 2 (2020). https://doi.org/10.3389/fclim.2020.586199.

[8] H. Thunman, T. Berdugo Vilches, M. Seemann, J. Maric, I.C. Vela, S. Pissot, H.N.T. Nguyen, Circular use of plastics-transformation of existing petrochemical clusters into thermochemical recycling plants with 100% plastics recovery, Sustainable Materials and Technologies 22 (2019) e00124. https://doi.org/10.1016/j.susmat.2019.e00124.

[9] E. Commission, J.R. Centre, P. Garcia-Gutierrez, A. Amadei, D. Klenert, S. Nessi, D. Tonini, D. Tosches, F. Ardente, H. Saveyn, Environmental and economic assessment of plastic waste recycling – A comparison of mechanical, physical, chemical recycling and energy recovery of plastic waste, Publications Office of the European Union, 2023. https://doi.org/doi/10.2760/0472.

[10] P.P. Plehiers, S.H. Symoens, I. Amghizar, G.B. Marin, C. V. Stevens, K.M. Van Geem, Artificial Intelligence in Steam Cracking Modeling: A Deep Learning Algorithm for Detailed Effluent Prediction, Engineering 5 (2019) 1027–1040. https://doi.org/10.1016/j.eng.2019.02.013.

[11] G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, L. Yang, Physics-informed machine learning, Nature Reviews Physics 3 (2021) 422–440. https://doi.org/10.1038/s42254-021-00314-5.

[12] E. Maria Tronci, A.R.J. Downey, A. Mehrjoo, P. Chowdhury, D. Coble, Physics-Informed Machine Learning Part I: Different Strategies to Incorporate Physics into Engineering Problems, in: Data Science in Engineering Vol. 10, River Publishers, New York, 2025: pp. 1–6. https://doi.org/10.1007/978-3-031-68142-4_1.

[13] J.D. Toscano, V. Oommen, A.J. Varghese, Z. Zou, N. Ahmadi Daryakenari, C. Wu, G.E. Karniadakis, From PINNs to PIKANs: recent advances in physics-informed machine learning,

Machine Learning for Computational Science and Engineering 1 (2025) 15. https://doi.org/10.1007/s44379-025-00015-1.

[14] D. Förtsch, K. Pabst, E. Groß-Hardt, The product distribution in Fischer–Tropsch synthesis: An extension of the ASF model to describe common deviations, Chem. Eng. Sci. 138 (2015) 333–346. https://doi.org/10.1016/j.ces.2015.07.005.

[15] D. Zhao, X. Wang, J.B. Miller, G.W. Huber, The Chemistry and Kinetics of Polyethylene Pyrolysis: A Process to Produce Fuels and Chemicals, ChemSusChem 13 (2020) 1764–1774. https://doi.org/10.1002/cssc.201903434.

[16] K.M. Van Geem, M.F. Reyniers, G.B. Marin, Challenges of Modeling Steam Cracking of Heavy Feedstocks, Oil & Gas Science and Technology - Revue de l'IFP 63 (2008) 79–94. https://doi.org/10.2516/ogst:2007084.

[17] F. Dai, H. Wang, M. Gong, C. Li, Z. Li, S. Zhang, Carbon-Number-Based Kinetics, Reactor Modeling, and Process Simulation for Coal Tar Hydrogenation, Energy & Fuels 29 (2015) 7532–7541. https://doi.org/10.1021/acs.energyfuels.5b01476.

[18] S. Shi, W. Tan, J. Sun, Progress in kinetic predictions for complex reaction of hydrocarbons: from mechanism studies to industrial applications, Reviews in Chemical Engineering 0 (2016). https://doi.org/10.1515/revce-2015-0029.

[19] F. Dai, Y. Yang, H. Wang, C. Li, Z. Li, S. Zhang, Pure carbon-number components to characterize the hydrocarbon mixture for kinetic modeling of hydrogenation process, Fuel 202 (2017) 287–295. https://doi.org/10.1016/j.fuel.2017.03.010.

[20] X. Zhang, K.K. Yalamanchi, S.M. Sarathy, A functional-group-based approach to modeling real-fuel combustion chemistry – I: Prediction of stoichiometric parameters for lumped pyrolysis reactions, Combust. Flame 227 (2021) 497–509. https://doi.org/10.1016/j.combustflame.2020.10.038.

[21] X. Zhang, S.M. Sarathy, A functional-group-based approach to modeling real-fuel combustion chemistry – II: Kinetic model construction and validation, Combust. Flame 227 (2021) 510–525. https://doi.org/10.1016/j.combustflame.2020.10.039.

[22] H. Thunman, F. Niklasson, F. Johnsson, B. Leckner, Composition of Volatile Gases and Thermochemical Properties of Wood for Modeling of Fixed or Fluidized Beds, Energy & Fuels 15 (2001) 1488–1497. https://doi.org/10.1021/ef010097q.

[23] D. Neves, H. Thunman, A. Matos, L. Tarelho, A. Gómez-Barea, Characterization and prediction of biomass pyrolysis products, Prog. Energy Combust. Sci. 37 (2011) 611–630. https://doi.org/10.1016/j.pecs.2011.01.001.

[24] F. Kirnbauer, M. Koch, R. Koch, C. Aichernig, H. Hofbauer, Behavior of Inorganic Matter in a Dual Fluidized Steam Gasification Plant, Energy & Fuels 27 (2013) 3316–3331. https://doi.org/10.1021/ef400598h.

[25] S.J.. Russell, Peter. Norvig, Artificial intelligence: A Modern Approach, Pearson Education, 2009.

[26] R. Dechter, Constraint Optimization, in: Constraint Processing, Elsevier, 2003: pp. 363–397. https://doi.org/10.1016/B978-155860890-0/50014-1.

[27] Francesca. Rossi, Peter. Van Beek, Toby. Walsh, Handbook of constraint programming, Elsevier, 2007.

[28] R. Dechter, Constraint Networks, in: Constraint Processing, Elsevier, 2003: pp. 25–49. https://doi.org/10.1016/B978-155860890-0/50003-7.

[29] F. Bacchus, X. Chen, P. van Beek, T. Walsh, Binary vs. non-binary constraints, Artif. Intell. 140 (2002) 1–37. https://doi.org/10.1016/S0004-3702(02)00210-2.

[30] Rina. Dechter, Constraint processing, Morgan Kaufmann Publishers, 2011.

[31] V. Kumar, Algorithms for Constraint-Satisfaction Problems: A Survey, AI Mag. 13 (1992) 32. https://doi.org/10.1609/aimag.v13i1.976.

[32] B. Selman, H.A. Kautz, B. Cohen, Noise Strategies for Improving Local Search, in: AAAI Conference on Artificial Intelligence, 1994. https://api.semanticscholar.org/CorpusID:6950786.

[33] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by Simulated Annealing, Science (1979). 220 (1983) 671–680. https://doi.org/10.1126/science.220.4598.671.

[34] A.E. Olsson, Particle Swarm Optimization: Theory, Techniques and Applications, Nova Science Publishers, 2011. https://books.google.se/books?id=Iad6RAAACAAJ.

[35] J. Sannemo, Principles of Algorithmic Problem Solving, 2018. http://csc.kth.se/~jsannemo/slask/main.pdf.

[36] C. Meng, S. Griesemer, D. Cao, S. Seo, Y. Liu, When physics meets machine learning: a survey of physics-informed machine learning, Machine Learning for Computational Science and Engineering 1 (2025) 20. https://doi.org/10.1007/s44379-025-00016-0.

[37] I. Florescu, Probability and stochastic processes, Wiley, 2015.

[38] J. Hioe, H. Zipse, Radical Stability—Thermochemical Aspects, in: Encyclopedia of Radicals in Chemistry, Biology and Materials, Wiley, 2012. https://doi.org/10.1002/9781119953678.rad012.

[39] M.L. Poutsma, The Radical Stabilization Energy of a Substituted Carbon-Centered Free Radical Depends on Both the Functionality of the Substituent and the Ordinality of the Radical, J. Org. Chem. 76 (2011) 270–276. https://doi.org/10.1021/jo102097n.

[40] Y.-R. Luo, Handbook of Bond Dissociation Energies in Organic Compounds, CRC Press, 2002. https://doi.org/10.1201/9781420039863.

[41] D. Neves, H. Thunman, A. Matos, L. Tarelho, A. Gómez-Barea, Characterization and prediction of biomass pyrolysis products, Prog. Energy Combust. Sci. 37 (2011) 611–630. https://doi.org/10.1016/j.pecs.2011.01.001.

[42] D. Xu, X. Zhang, Z. Chen, L. Zhang, Molecular-Level Kinetic Model for Light Hydrocarbon Steam Cracking Based on the SU-BEM Framework, ACS Omega 9 (2024) 14849–14859. https://doi.org/10.1021/acsomega.3c07749.

[43] S.E. Levine, L.J. Broadbelt, Detailed mechanistic modeling of high-density polyethylene pyrolysis: Low molecular weight product evolution, Polym. Degrad. Stab. 94 (2009) 810–822. https://doi.org/10.1016/J.POLYMDEGRADSTAB.2009.01.031.

[44] Z. Gholami, F. Gholami, Z. Tišler, M. Vakili, A Review on the Production of Light Olefins Using Steam Cracking of Hydrocarbons, Energies (Basel). 14 (2021) 8190. https://doi.org/10.3390/en14238190.

[45] M. Altarawneh, L. Ali, Formation of Polycyclic Aromatic Hydrocarbons (PAHs) in Thermal Systems: A Comprehensive Mechanistic Review, Energy & Fuels 38 (2024) 21735–21792. https://doi.org/10.1021/acs.energyfuels.4c03513.

[46] E. Reizer, B. Viskolcz, B. Fiser, Formation and growth mechanisms of polycyclic aromatic hydrocarbons: A mini-review, Chemosphere 291 (2022) 132793. https://doi.org/10.1016/j.chemosphere.2021.132793.

[47] M. Frenklach, A.M. Mebel, On the mechanism of soot nucleation, Physical Chemistry Chemical Physics 22 (2020) 5314–5331. https://doi.org/10.1039/D0CP00116C.

[48] T. Chen, T. Li, J. Sjöblom, H. Ström, A reactor-scale CFD model of soot formation during high-temperature pyrolysis and gasification of biomass, Fuel 303 (2021) 121240. https://doi.org/10.1016/j.fuel.2021.121240.

[49] I. T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 2002. https://doi.org/10.1007/b98835.

[50] J. Weitkamp, Catalytic Hydrocracking—Mechanisms and Versatility of the Process, ChemCatChem 4 (2012) 292–306. https://doi.org/10.1002/cctc.201100315.

[51] A. de Klerk, Thermal Cracking of Fischer−Tropsch Waxes, Ind. Eng. Chem. Res. 46 (2007) 5516–5521. https://doi.org/10.1021/ie070155g.

[52] S. Chakraborty, Generating discrete analogues of continuous probability distributions-A survey of methods and constructions, J. Stat. Distrib. Appl. 2 (2015) 6. https://doi.org/10.1186/s40488-015-0028-6.

[53] S. Coles, An Introduction to Statistical Modeling of Extreme Values, Springer London, London, 2001. https://doi.org/10.1007/978-1-4471-3675-0.

[54] T. Berdugo Vilches, J. Marinkovic, M. Seemann, H. Thunman, Comparing Active Bed Materials in a Dual Fluidized Bed Biomass Gasifier: Olivine, Bauxite, Quartz-Sand, and Ilmenite, Energy & Fuels 30 (2016) 4848–4857. https://doi.org/10.1021/acs.energyfuels.6b00327.

[55] M. Israelsson, A. Larsson, H. Thunman, Online Measurement of Elemental Yields, Oxygen Transport, Condensable Compounds, and Heating Values in Gasification Systems, Energy & Fuels 28 (2014) 5892–5901. https://doi.org/10.1021/ef501433n.

[56] W. Kaminsky, M. Predel, A. Sadiki, Feedstock recycling of polymers by pyrolysis in a fluidised bed, Polym. Degrad. Stab. 85 (2004) 1045–1050. https://doi.org/10.1016/j.polymdegradstab.2003.05.002.

[57] M.S. Abbas-Abadi, The effect of process and structural parameters on the stability, thermo-mechanical and thermal degradation of polymers with hydrocarbon skeleton containing PE, PP, PS, PVC, NR, PBR and SBR, J. Therm. Anal. Calorim. 143 (2021) 2867–2882. https://doi.org/10.1007/s10973-020-09344-0.

[58] Y. Liu, J. Qian, J. Wang, Pyrolysis of polystyrene waste in a fluidized-bed reactor to obtain styrene monomer and gasoline fraction, Fuel Processing Technology 63 (2000) 45–55. https://doi.org/10.1016/S0378-3820(99)00066-1.

[59] I. Maafa, Pyrolysis of Polystyrene Waste: A Review, Polymers (Basel). 13 (2021) 225. https://doi.org/10.3390/polym13020225.

[60] M. Herrera, G. Matuschek, A. Kettrup, Main products and kinetics of the thermal degradation of polyamides, Chemosphere 42 (2001) 601–607. https://doi.org/10.1016/S0045-6535(00)00233-2.

[61] M. Nielsen, P. Jurasek, J. Hayashi, E. Furimsky, Formation of toxic gases during pyrolysis of polyacrylonitrile and nylons, J. Anal. Appl. Pyrolysis 35 (1995) 43–51. https://doi.org/10.1016/0165-2370(95)00898-O.

[62] J. Sun, T. Ding, X. Zhao, P. Wang, W. Chen, J. Yu, Dehydrochlorination of PVC at low-temperatures and the pyrolysis behavior of dechlorinated PVC, Waste Dispos. Sustain. Energy (2025). https://doi.org/10.1007/s42768-025-00252-3.

[63] H. Zhou, C. Wu, J.A. Onwudili, A. Meng, Y. Zhang, P.T. Williams, Influence of process conditions on the formation of 2–4 ring polycyclic aromatic hydrocarbons from the pyrolysis of polyvinyl chloride, Fuel Processing Technology 144 (2016) 299–304. https://doi.org/10.1016/J.FUPROC.2016.01.013.

[64] G.F.. Lawler, Introduction to stochastic processes, Chapman & Hall/CRC, 2006.

[65] C.M.. Grinstead, Introduction to probability, Amer Mathematical Soc, 2012.

[66] C. Mandviwala, Steam Cracking of Polyolefins in Fluidized Bed Systems: Influences of the Bed Materials on the Hydrogen Transfer Reactions, n.d.

[67] D.F. Anderson, T.G. Kurtz, Continuous Time Markov Chain Models for Chemical Reaction Networks, in: Design and Analysis of Biomolecular Circuits, Springer New York, New York, NY, 2011: pp. 3–42. https://doi.org/10.1007/978-1-4419-6766-4_1.

[68] J. Gillariose, J. Joseph, C. Chesneau, Lasso and Ridge regression: a comprehensive review of applications and developments in machine learning, Int. J. Data Sci. Anal. 21 (2026) 7. https://doi.org/10.1007/s41060-025-00957-y.

[69] A.J. Dobson, A.G. Barnett, An Introduction to Generalized Linear Models, CRC Press, 2018. https://books.google.se/books?id=AS_3DwAAQBAJ.

[70] Gabriel Furnieles, https://towardsdatascience.com/sigmoid-and-softmax-functions-in-5-minutes-f516c80ea1f9/, (2022). https://towardsdatascience.com/sigmoid-and-softmax-functions-in-5-minutes-f516c80ea1f9/ (accessed November 3, 2025).

[71] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, (2014). https://arxiv.org/abs/1412.6980.

[72] ADAM PyTorch Implementation, WebPage (n.d.). https://docs.pytorch.org/docs/stable/generated/torch.optim.Adam.html (accessed October 17, 2025).

[73] ADAM TensorFlow Implementation, WebPage (n.d.). https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam (accessed October 17, 2025).

[74] T. Berdugo Vilches, H. Thunman, Experimental Investigation of Volatiles–Bed Contact in a 2–4 MWth Bubbling Bed Reactor of a Dual Fluidized Bed Gasifier, Energy & Fuels 29 (2015) 6456–6464. https://doi.org/10.1021/acs.energyfuels.5b01303.

[75] M. Israelsson, M. Seemann, H. Thunman, Assessment of the Solid-Phase Adsorption Method for Sampling Biomass-Derived Tar in Industrial Environments, Energy & Fuels 27 (2013) 7569–7578. https://doi.org/10.1021/ef401893j.

[76] A. Larsson, M. Seemann, D. Neves, H. Thunman, Evaluation of Performance of Industrial-Scale Dual Fluidized Bed Gasifiers Using the Chalmers 2–4-MW $_{th}$ Gasifier, Energy & Fuels 27 (2013) 6665–6680. https://doi.org/10.1021/ef400981j.

[77] SCS — SCS 3.2.1 documentation, (n.d.). https://www.cvxgrp.org/scs/ (accessed October 12, 2022).

[78] S. Diamond, S. Boyd, CVXPY: A Python-Embedded Modeling Language for Convex Optimization, J. Mach. Learn. Res. 17 (2016) 2909–2913.

NOMENCLATURE

ADAM: Adaptive Moment Estimation (gradient-based optimizer)

C-AL: Carbon bonds to Aliphatics

C-AR: Carbon bonds to Aromatics

C-xO: Carbon bonds to Oxygen

BDE: Bond Dissociation Energy

CBG: Carbon Bond Group model

CCU: Carbon Capture and Utilization

CN: Constraint Network

COP: Constraint Optimization Problem

CSP: Constraint Satisfaction Problem

CVXPY: Python domain-specific language for convex optimization

DFB: Dual Fluidized Bed reactor

DSL: Domain-Specific Language

GC–VUV: Gas Chromatography – Vacuum Ultraviolet Spectroscopy

GLM: Generalized Linear Model

LHV: Lower Heating Value

MAE: Mean Absolute Error

NN: Neural Network

OSQP: Operator Splitting Quadratic Program (solver)

PI: Physics-Informed

PICN: Physics-Informed Constraint Network

PINN: Physics-Informed Neural Network

PSM: Parametric System Model

Q, R, N, B, K: Markov transition matrices (transient, absorbing, and mapping)

R^2: Coefficient of Determination

$\mathbb{R}$: Real Numbers ($\mathbb{R}^+$: Positive Real numbers, $\mathbb{R}^3$: the set of all triples of real numbers)

RMSE: Root Mean Square Error

SC: Steam Cracking

SCS: Splitting Conic Solver

SPA: Solid Phase Adsorption method

Δb: Measurement uncertainty bound

Δn: Change in number of moles

$\phi(u)$: Feature mapping function

λ: Regularization coefficient / trade-off parameter

I: Identity matrix

APPENDIX (Definitions of Technical Terms)

❖ Chemical & Process Engineering

Bond Dissociation Energy (BDE): The energy required to break a specific chemical bond homolytically; it quantifies the bond's strength and influences where and how polymers crack under heat.

Carbon Capture and Utilization (CCU): A technology that captures CO_2 from combustion or process emissions and reuses it as a carbon source for fuels or chemicals, closing the carbon loop.

Dual Fluidized Bed (DFB): A reactor composed of two interconnected fluidized beds — one for combustion (heat generation) and one for endothermic conversion — allowing indirect heat transfer for stable, oxygen-free reactions such as pyrolysis or steam cracking.

Endothermic process: A reaction that requires heat input to proceed, common in cracking and pyrolysis reactions.

Free radical reactions: Chain reactions initiated by thermal bond cleavage that produce reactive species (radicals) which propagate decomposition and recombination of hydrocarbons during cracking.

Producer gas: The raw gas mixture exiting the reactor, containing syngas and hydrocarbons; its composition depends on feedstock and operating conditions.

Steam cracking: A high-temperature process (700–850 °C) in which hydrocarbons or polymers are broken down in steam into smaller molecules like ethylene, propylene, and aromatics. Used for both monomer production and chemical recycling.

Syngas: A gaseous mixture of CO and H_2 produced during thermochemical conversion, used as a precursor for fuel or chemical synthesis.

Thermochemical conversion: A broad term for chemical transformations driven by heat, including pyrolysis, gasification, and combustion, which convert solid or liquid carbonaceous feedstocks into gases, oils, and solids.

❖ Modelling & Mathematics

Absorbing Markov chain: A stochastic model where some states (like final product groups) are absorbing, meaning once entered they cannot change; used to represent irreversible conversions.

Barycentric coordinates: Weights at the triangle's vertices that sum to one; represent any interior composition point.

Carbon-Bond-Group (CBG) Model: A framework grouping carbon atoms by their bonding environment — to heteroatoms (C–X), aromatics (C–AR), or aliphatics (C–AL) — allowing direct comparison between feed and product structures.

Column-stochastic K mapping: K maps feed simplex to product simplex ensuring non-negativity and column sums equal to one.

Column-stochastic matrix: Matrix with non-negative entries whose columns each sum to one; maps one simplex (input composition) to another (output products).

Compositional vector: Vector of component shares constrained to be non-negative and sum to one; used for feed/product bond group compositions.

Constraint Network (CN): A mathematical structure of variables, domains, and constraints used to represent and solve problems under defined rules; forms the theoretical basis for the models in this work.

Constraint Optimization Problem (COP): Similar to a CSP but allows trade-offs by minimizing a cost function when all constraints cannot be satisfied exactly.

Constraint Satisfaction Problem (CSP): A problem where variables must take values that satisfy all constraints simultaneously, e.g., finding a carbon balance that fulfills all conservation laws.

Convex combination: A convex combination is a weighted average of vectors with non-negative weights that sum to one, ensuring results remain within the feasible space. It underpins the mathematical operations that map feed compositions to product compositions in the CBG model.

Domain constraints / Inequality constraints: Bounds or relations that variables must satisfy (e.g., $0 \leq x \leq 1, |Ax - b| \leq \Delta b$); define feasible set around measurements.

Domain-related priors: Physics or Chemistry-based prior knowledge that can be added to the model as constraints (hard or soft penalties).

Euclidean norm (L2): Standard vector length; used in squared-error losses and ridge regularization penalties.

Frobenius norm: Matrix analog of Euclidean norm; used to penalize matrix parameters in regularization.

Hard constraint: A rule that must always be fulfilled, such as mass or energy conservation

Invariant: A quantity that remains constant despite changes in system conditions (e.g., total carbon mass), used to ensure physical consistency.

Jacobian / Kronecker delta: Derivatives of SoftMax with respect to logits use the Kronecker delta; Jacobian helps compute gradients for A and B parameters.

Markov chain (absorbing): Stochastic process with transient and absorbing states; absorption matrix maps initial bond groups to final product groups.

Mass / Elemental conservation: A fundamental constraint ensuring the number of atoms or total mass is constant before and after conversion; implemented as a hard constraint.

Monovariant: A property that changes in only one direction (e.g., entropy increase), used to check model stability.

Nonnegativity constraint: Requires variables or matrix entries to be ≥0; enforces physical feasibility (e.g., yields/probabilities).

Objective / Loss function: A scalar function measuring fit or error; minimized subject to constraints to estimate parameters or states.

Parametric System Model (PSM): A compact mathematical framework where measured species data are expressed as parameterized distribution functions under mass and elemental constraints to evaluate data quality and estimate unknowns.

Probability distribution function: A mathematical function describing how likely different outcomes (e.g., molecule sizes) are; used in the PSM to represent carbon-group distributions.

Pseudo-bond channel: Auxiliary 'reactor fingerprint' pathway that redistributes composition toward reactor-specific tendencies without adding mass.

Ridge (L2) regularization: Penalty on parameter magnitude to prevent overfitting and improve generalization.

Simplex / Ternary simplex: A simplex is a geometric representation of all possible combinations of a set of components that sum to one, such as compositional vectors. In this work, ternary simplices illustrate three-component carbon-bond compositions, and simplex mappings ensure physically valid (non-negative, normalized) conversions.

Soft constraint: A preference or penalty condition that can be slightly violated at a cost, often used to fine-tune model fitting.

Topology constraint: A rule derived from function shape or convergence behavior (e.g., ensuring distribution tails decay realistically).

Transition / Absorption matrices (Q, R, N, B): Q: transient transitions; R: transient to absorbing transitions; $N = (I - Q)^{-1}$ is the fundamental matrix; $B = NR$ gives absorption probabilities.

Transition matrix: A matrix describing the probabilities of moving between states (e.g., bond groups) in a system; used to model conversion pathways.

❖ Data Science & Machine Learning

ADAM (Adaptive Moment Estimation): is a stochastic gradient-based optimizer (see gradient descent) that adjusts learning rates adaptively using running averages of first and second moments of the gradients. It combines the advantages of other gradient methods (AdaGrad and RMSProp) and is robust for noisy, high-dimensional data. It is widely used for training neural networks and nonlinear regression models in constrained optimization settings.

Adaptive control: A control strategy where the model continuously updates its parameters in response to process feedback for stability and optimization.

Convex Optimization: A class of optimization problems where both the objective function and feasible region are convex, guaranteeing that any local minimum is also a global minimum. It provides stable and efficient solutions for CN problems of this type.

CVXPY (DSL): Python domain-specific language to formulate and solve convex optimization problems with multiple solvers.

Descriptive, Predictive, Prescriptive analytics: Three levels of data analysis: describing existing data, forecasting future trends, and recommending optimal actions based on models.

Dimensionality reduction: Mathematical methods (e.g., PCA) that simplify complex datasets by compressing correlated variables into fewer informative dimensions.

Feature map $\phi(u)$: Vector of engineered condition features (e.g., T, S/C, interactions, splines) feeding the logits for condition-dependent mappings.

First-order solvers (SCS, OSQP): Convex optimization algorithms solving large constrained problems efficiently; accessible via CVXPY DSL.

Generalization (Model Property): The ability of a model trained on one dataset to accurately predict unseen cases; essential for applying lab data to other reactors.

Generalized Linear Model (GLM): Statistical model linking expected response to a linear predictor via a link function (e.g., SoftMax for multinomial outcomes).

Grid search / Hyperparameters / Metric: Systematic exploration of algorithm settings (e.g., solvers, losses, polymers set) evaluated by a chosen error metric.

Gradient Descent: An iterative algorithm that updates parameters by moving in the direction of the negative gradient to reduce the loss function. It is widely used for training models in machine learning and optimization.

Hierarchical (multi-reactor) parameters: Parameterization with shared global components and reactor-specific deviations; regularization controls overfitting.

Learning rate / Epoch: Learning rate controls step size in optimization; an epoch is one full pass over the training data.

Linear/Non-Linear Solvers: Algorithms that compute solutions to systems of equations: linear solvers handle relationships of the form $Ax = b$, while non-linear solvers address more complex, non-linear relationships through iterative methods. They are commonly used to solve balance equations and parameter estimation problems.

Logits: Log-odds transformation used in statistical and machine learning models to map probabilities into the real-number domain. In this work, logits represent unconstrained parameters before applying the SoftMax function to obtain normalized compositional probabilities in the CBG mapping.

Metaheuristic Method: A high-level optimization strategy (e.g., PSO, simulated annealing) that explores large and complex solution spaces using stochastic or adaptive rules. It does not guarantee a global optimum as in convex optimization but often finds good approximate solutions when deterministic methods struggle.

Monte Carlo Method: A numerical technique that uses repeated random sampling to estimate quantities such as distributions, expectations, or uncertainties. It is commonly applied when analytical or deterministic solutions are infeasible.

Physics-Informed Model / PICN: Model or constraint network embedding physical laws/invariants as priors into data-driven algorithms to limit solutions to physically meaningful regions, ensuring predictions obey known principles.

Physics-Informed Neural Network (PINN): A neural network trained with both data and physical laws (encoded as differential or algebraic constraints) to improve interpretability and reduce data needs.

PSO (Particle Swarm Optimization): Metaheuristic optimizer exploring solution space via a population of candidate solutions; alternative to gradient methods.

Regularization: A mathematical penalty added to machine learning models to prevent overfitting and enforce smoothness or physical plausibility.

Ridge Regularization: A regularization technique (L2 penalty) that discourages large parameter values by adding the squared magnitude of the weights to the loss function. It improves model stability and prevents overfitting in regression and machine-learning models.

Sigmoid: S-shaped function mapping real numbers to (0,1); used to bound a mixing weight (e.g., pseudo-bond channel).

SoftMax: Mathematica transformation that converts unconstrained logit scores into probabilities (positive, sum-to-one), keeping K column-stochastic.

Torch / Tensor operations: PyTorch library and array operations used to implement and accelerate gradient-based training.